STATISTICAL MODELLING

Proceedings of the 19th International Workshop on Statistical Modelling

Florence (Italy) 4-8 July, 2004

Annibale Biggeri
Emanuela Dreassi
Corrado Lagazio
Marco Marchi
(eds)

Dipartimento di Statistica
"Giuseppe Parenti"

Firenze University Press

Annibale Biggeri, Emanuela Dreassi,
Corrado Lagazio, Marco Marchi (Eds)

Statistical Modelling

Proceedings of the
19th International Workshop on Statistical Modelling
Florence (Italy) 4-8 July, 2004

Editors:

Annibale Biggeri
Dip. di Statistica "G. Parenti"
Università degli Studi di Firenze
Viale Morgagni, 59
I 50134 Florence (Italy)

Emanuela Dreassi
Dip. di Statistica "G. Parenti"
Università degli Studi di Firenze
Viale Morgagni, 59
I 50134 Florence (Italy)

Corrado Lagazio
Dip. di Scienze Statistiche
Università di Udine
Via Treppo 18
I 33100 Udine (Italy)

Marco Marchi
Dip. di Statistica "G. Parenti"
Università degli Studi di Firenze
Viale Morgagni, 59
I 50134 Florence (Italy)

Preface

This volume collects the proceedings of the 19th International Workshop on Statistical Modelling, held in Florence, Italy, 4 to 8 July, 2004.

The International Workshop on Statistical Modelling has been held in Europe and the USA for the past twenty years. The workshop arose out of two GLIM conferences in the U.K. in London (1982) and Lancaster (1985), and focused on various aspects of statistical modelling in an informal environment, specifically aimed at applied statistics, but also including theoretical developments and computational methods. The spirit of the workshop has always concentrated on papers that are motivated by real life data and make novel contributions to the subject. Statistical modelling is an important cornerstone in many scientific disciplines, and the workshop has consistently provided a rich environment for cross-fertilization of ideas from different statistical disciplines. The workshop has brought together scientists from different nationalities with different backgrounds and experience, and has thus always promoted contributions from students early in their careers and allowed time for discussion and interchange between junior and senior scientists. The inaugural workshop in this series took place in Innsbruck in 1986, and since then the workshop has grown substantially, and now regularly attracts over 150 participants. There has been a strong effort made to bring each new meeting to a different European country: Perugia (1987), Vienna (1988), Trento (1989), Toulouse (1990), Utrecht (1991), Munich (1992), Leuven (1993), Exeter (1994), Innsbruck (1995), Orvieto (1996), Biel/Bienne (1997) - to the USA - New Orleans (1998) - and back to Europe - Graz (1999), Bilbao (2000), Odense (2001), Chania (2002), Leuven (2003), Florence (2004). The year 2005 will take the workshop to Australia.

The Florence workshop consists in 48 oral presentations and 47 posters; four invited lectures complete the lay-out. The oral contributions are arranged in eight sessions: *Statistical Modelling in Genomics and Genetics* will offer a broad perspective on this new field of applied research, with Geoff MacLachlan, Avner Bar-Hen, Ernst Wit, Ib M. Skovgaard, among others from the Italian group recently funded by the Ministry of Education and Research. *Semi-parametric Regression Models* presents important new research ideas with Paul Eilers, Vicente Núñez-Antón, Marc Saez, and interesting student presentations from the groups of Göran Kauermann and

Peter Diggle. However the main focus of the workshop is on *Generalized Linear Mixed Models*. Papers on this topic will also be presented in several other sessions: *Correlated Data Modelling*, *Missing Data*, *Measurements Error* and *Survival Analysis*. A great deal of important results relevant on statistical modelling will be discussed, by the group of Emmanuel Lesaffre, and of Geert Molenberghs, by Peter van der Heijden, Regina Tüchler, George Streftaris, Brent Coull and many others I cannot quote here. Finally two specialized sessions on *Spatial Data Modelling* (with Renato Assunção presenting new research ideas) and *Time Series and Econometrics* (with interesting student presentations) will complete the workshop. Many Italian researchers will attend the workshop and the list is so good and long that it prevents me from quoting someone in particular.

We had a very difficult task in selecting oral presentations from the over one hundred submissions. Therefore many good papers were reported in the Poster Session. The reader and the attender is recommended to pay careful attention to this not secondary part of the workshop.

I wish to conclude mentioning the four invited speakers. First we were able to organize a special event under the sponsorship of the Nutrigenomics Organization (NuGO), a European VI Framework funded research network. Terry Speed will speak on "Statistical analysis of replicated microarray time series data", which is on the focus of many study designs in Functional Genomics nowadays. Terry Speed is an outstanding personality and contributed to systematize and clarify the statistical issues in the analysis of gene expression data. Generalized Linear Mixed Models and their link to Latent variables modelling will be stressed by the second invited speaker, Anders Skrondal, jointly with Sophia Rabe-Hesketh. Advances in computational issues and a very general theoretical frame will be introduced, which makes their contribution one of the most stimulating for applied statisticians. Stuart Coles will discuss on "A censored point process model for extreme volcanic eruptions" and his lecture will highlight the subtleties and potentiality of statistical modelling where theory and sensibility to subject-specific issues create the special flavour and appeal of applied statistics. Roberto Colombi will speak on "Marginal models: recent developments and applications to categorical time series analysis". This is a classical topic in correlated data modelling consistent with the tradition of our workshop. Roberto Colombi's paper offers a review and new methodological insights in such area, still one of the most popular among applied statisticians.

The Editors of this volume would like to thank all members of the Scientific Committee and other referees who worked hard in assessing the papers submitted. The local organisers of the workshop listed below also deserve our gratitude. We are very thankful to the authors who have considerably simplified the task of preparing this volume by submitting their papers in

LaTeX and by keeping strictly to the tight timescale.

A special thank finally to Cristina Dolfi and Marie-Hélène Piette for their valuable contribution to the scientific and organising secretariat, and to the webmaster Nicola Nostro. The workshop owes its final shape and its success mostly to their intelligence and application.

Florence, May 28, 2004

Annibale Biggeri

Scientific Programme Committee: Adelchi Azzalini (Padova), Avner Bar-Hen (Marseille), Adrian Bowman (Glasgow), Antonio Forcina (Perugia), Arnoldo Frigessi (Oslo), Dominique Guguan (Cachan), Leonhard Held (Munich), Brunero Liseo (Roma), Giovanni M. Marchetti (Firenze), Kenan M. Matawie (Sydney), Vicente Núñez-Antón (Bilbao), Gianpaolo Scalia Tomba (Roma), Gilg Seeber (Innsbruck), Terry Speed (Berkeley), Gordon Smyth (Victoria), Bill Venables (Cleveland)

Local Organising Committee: Annibale Biggeri (Firenze), Monica Chiogna (Padova), Mauro Gasparini (Torino), Corrado Lagazio (Udine), Marco Marchi (Firenze)

Contents

Index of Authors

Invited Sessions

A Censored Point Process Model for Extreme Volcanic Eruptions

Stuart Coles[1]

[1] Dipartimento di Scienze Statistiche, Via C. Battisti 214/243, 35121 Padova, Italia.

Abstract: The magnitude of a volcanic eruption is an essential component of risk assessment in volcano-sensitive regions. Extreme value models are a natural candidate for modelling such phenomena, and a point process representation for extreme value behaviour provides a convenient inferential framework. However, direct application to databases of volcanic events is complicated by an under-recording of historical events. This is complicated further by the fact that small events appear to have a greater tendency to go unreported relative to large events. In this article we suggest modifying the standard point process model for extremes with a parametric component that models the under-reporting mechanism.

Keywords: Extreme values, Point processes, Volcanoes.

1 Introduction

Let me come clean: I originally prepared a version of this article for presentation at a workshop on Statistics in Volcanology, held at Bristol University. Volcanological models are traditionally deterministic, and statistics in this field is generally used to mop-up noise when real-life observations turn out to be different from model predictions. I wanted to give a presentation that emphasised the benefits of developing models that incorporated a stochastic element. My idea was really just to invent a problem, based on some volcanological data that I was able to track down on the web, and to develop a hypothetical model that would serve as a metaphor for the possible integration of scientific knowledge into a statistical model. The particular model is partly motivated by representations for extreme value behaviour, and partly by an understanding of volcanic processes. However, some parts of the model are rather arbitrary and open to improvement. As it turned out, the volcanolgists at the workshop were enthused by the analysis itself. It remains to be seen if they also take on board the wider methodological issues I was trying to propose. This article summarises the ideas.
Volcanology is an essential science, partly for a geological understanding of the earth's composition, but more crucially to enable a calculation of hazard risk in regions prone to volcanic activity. In such regions, various

criteria need to be taken into account in civil protection schemes: the morphology that determines direction of lava flow, the likelihood of a future eruption, and the plausible values of an eruption magnitude. More generally, since volcanic eruptions are potentially the most explosive naturally occurring processes on Earth, there is genuine scientific interest in quantifying a worst-case scenario for future events (Mason *et al.*, 2004). Though these questions are undoubtedly difficult to address statistically, their nature suggests that extreme value theory might provide a more plausible class of models to work with than would other areas of statistics.
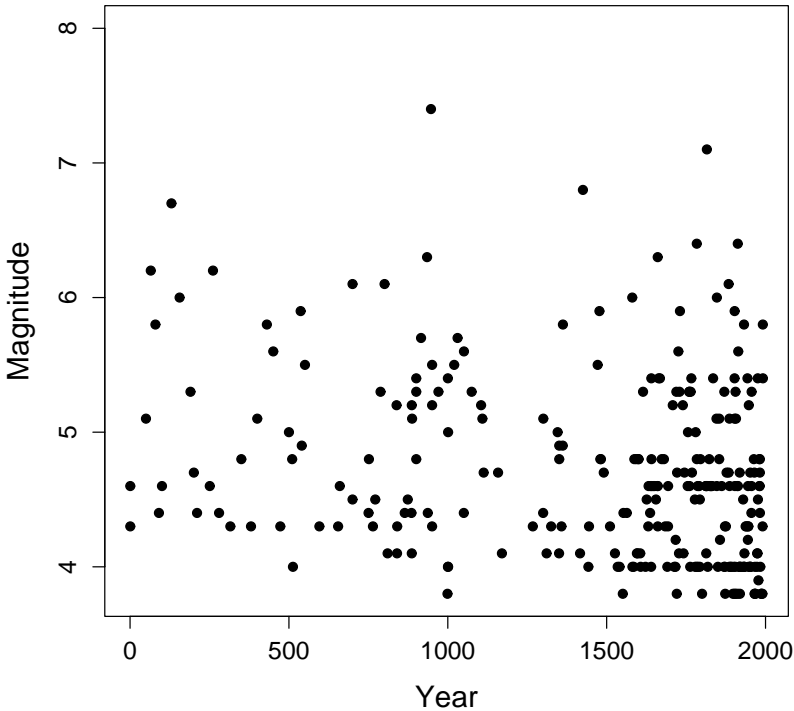
FIGURE 1. Historical catalogue of volcanic eruptions with magnitudes exceeding $x = 3.7$.

There are different definitions of the magnitude of a volcanic eruption, but one commonly used version is in terms of the mass of magma generated; specifically $X = \log m - 7$, where m is the mass of magma released during

the eruption in kg. Volcanologists have a variety of means to approximate the value of X, even for volcanic events that are centuries old, though clearly the reliability of such measurements decreases with age. A plot of the available data is shown in Fig. 1. These data derive from a database that purports to include all known volcanic eruptions in the last 2 millennia, having a magnitude greater than $x = 3.7$ (Simkin and Siebert, 1994). A few additional events with unspecified magnitude have been dropped from both the figure and our subsequent analyses, though in principle such censored information could also be exploited.

One feature is strikingly evident from the Fig. 1: the rate of activity in the last 500 years or so is very much greater than in the previous 1500 years. But this is at odds with known volcanology, which suggests the rate of activity has been more or less constant over the period. There is also some suggestion in the figure that the rate of weaker volcanic events has changed more drastically than that of larger ones. Of course, a more realistic explanation for this phenomenon is that volcanic activity has remained constant over the period, but that historical events are harder to identify than recent ones, particularly if they are weak in magnitude. Consequently, the data in Fig. 1 are the result of two processes: the volcanic activity itself, followed by the recording process. Ignoring the measurement aspect of the problem could lead to potential bias in the assessment of the volcanic aspect.

2   Extreme values via point processes

There are different characterizations of the extremal properties of stochastic processes. One particularly convenient representation – both for theoretical treatment and modelling – is in terms of point processes. The theory for this approach is due to Pickands (1971), while Smith (1989) was the first to propose inference explicitly in this framework. In simple terms, suppose that $X_1, \ldots, X_n$ is a sequence of independent random variables with common distribution function F, and our interest is in modelling the tail of F. We define the point process $P_n = \{(i/(n+1), X_i) : i = 1, \ldots, n\}$. Under detailed limiting arguments that hold under very general conditions on F, it is reasonable to model the process P_n over the region $\mathcal{A}_u = [0, 1] \times [u, \infty)$, for a sufficiently large threshold u, as a non homogeneous Poisson process with intensity density function in the family

$$\lambda(t, x) = \frac{1}{\sigma} \left[1 + \xi \frac{(x - \mu)}{\sigma} \right]_+^{-1/\xi - 1}, \tag{1}$$

where $\sigma > 0$ and $a_+ = \max(a, 0)$. This is consistent, for example, with classical representations for extremes based on block maxima or threshold exceedances; see Coles (2001, ch.7) for a general discussion of these connections. Inference amounts to estimation of the parameters (μ, σ, ξ) on the

basis of the observed data in the region $\mathcal{A}_u$, $\{(t_1, x_1), \ldots, (t_m, x_m)\}$ say. The Poisson assumption immediately provides a likelihood function

$$L(\mu, \sigma, \xi; (t_1, x_1) \ldots, (t_n, x_n)) = n_y \exp\left\{-\int_{\mathcal{A}_u} \lambda(t, x) dt dx\right\} \prod_{i=1}^{n} \lambda(t_i, x_i),$$
(2)

where the inclusion of the proportionality constant n_y, defined as the number of years of observation, scales the parameterization of the model. The likelihood function (2) can then be used as the basis of either a classical or a Bayesian inference.
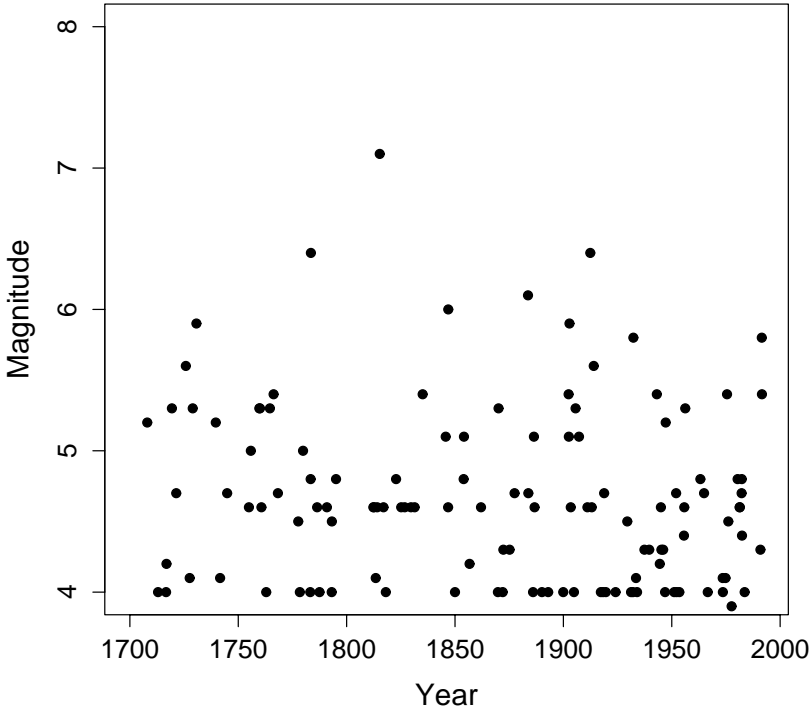
FIGURE 2. Recent volcanic eruptions with magnitudes exceeding $x = 3.7$.

For the volcano magnitude data, the assumed under-reporting of historical events implies that the assumption of time homogeneity is invalid. To illustrate the point process methodology though, we can restrict attention to

a more recent section of the data, which seems approximately stationary. Fig. 2 shows the events over the last 300 years or so.
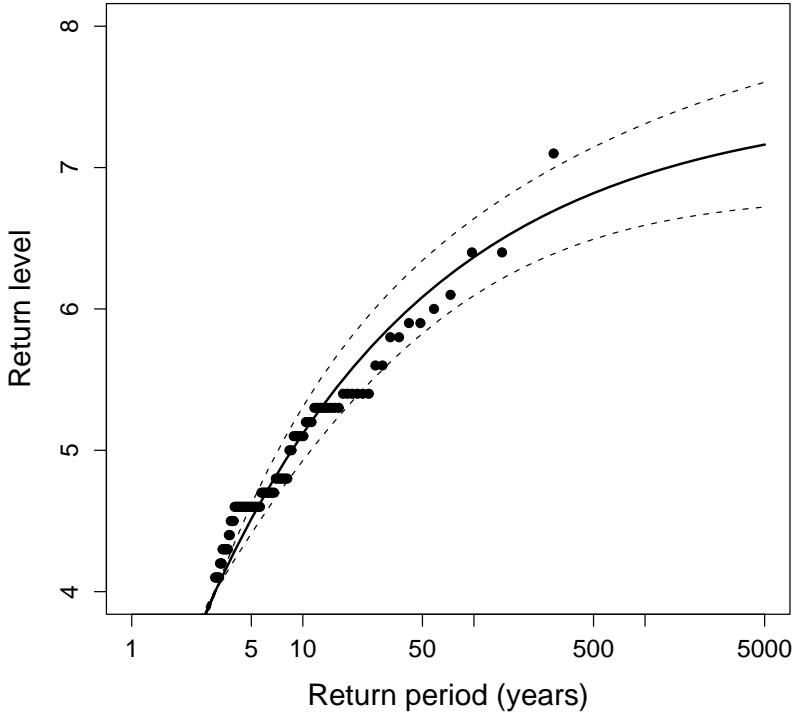
FIGURE 3. Return level plot of volcano magnitudes. Dashed curves show limits of 95% pointwise confidence intervals; points show empirical estimates.

There are a range of diagnostics to assist with threshold choice. A particularly simple diagnostic, the mean residual life plot (Davison and Smith, 1990), supports a threshold of $u = 4$ for these data. Based on this choice, the maximum likelihood estimates are obtained as $(\hat{\mu}, \hat{\sigma}, \hat{\xi}) = (2.45, 1.66, -0.330)$ with standard errors $(0.284, 0.297, 0.061)$ respectively. One important aspect of this result is the strong evidence for a negative value of ξ, implying a finite upper bound on volcanic eruption magnitudes. Other aspects are perhaps more easily interpreted after a transformation of results: the threshold exceedance rate (per year as a consequence of the scaling factor in (2)) is

given by

$$\eta = \frac{1}{\sigma}\left[1 + \xi\frac{(u-\mu)}{\sigma}\right]_+^{-1/\xi-1}, \tag{3}$$

and the conditional excess distribution by

$$P(X > x \mid X > u) = [1 + \xi\,(x-u)/\tilde{\sigma}]_+^{-1/\xi}, \tag{4}$$

where $\tilde{\sigma} = \sigma + \xi(u-\mu)$. Combining (3) and (4), it follows that the level $x > u$ is expected to be exceeded once every $r_p(x)$ years, where

$$r(x) = \eta^{-1}\left[1 + \xi\,(x-u)/\tilde{\sigma}\right]_+^{1/\xi}. \tag{5}$$

In common terminology, $r(x)$ is the return period associated with level x. Substitution of maximum likelihood estimates leads to an estimate of $\hat{\eta} = 0.28$ for η, and the return level plot (x against $r(x)$ on a logarithmic scale, as is common for such graphs) shown in Fig. 3.

3   A censored point process model

The point process set-up provides a convenient way to extend the analysis to allow for the under-recording of historical events as observed in Fig. 1. We assume that an event that occurred at time t and having magnitude x is actually recorded in the data catalogue with probability $p(t,x)$. Hence, the Poisson assumptions of the observed process are unchanged, except that the intensity function is modified to

$$\lambda_M(t,x) = p(t,x)\lambda(t,x). \tag{6}$$

This is the metaphor referred to in the opening paragraph. Without external knowledge, the data alone would be insufficient to formulate this model. However, knowing that the volcanic process has remained largely homogeneous in time, and understanding that under-reporting of historical events is a likely phenomenon that is plausibly more pronounced for weaker events, leads to the modified intensity model (6). This is the metaphor: $p(\cdot,\cdot)$ is formulated from scientific knowledge of the process; $\lambda(\cdot,\cdot)$ is determined from statistical considerations.

There are different ways forward at this point. In this article we take the approach of adopting a parametric family for $p(\cdot,\cdot)$ that conforms to our beliefs about the under-recording mechanism. Specifically, we choose a family for which:

1. $p(1,x) = 1$ for each x, corresponding to an assumption that any volcano with magnitude above the threshold level would be recorded at the present time;

2. $p(t, x)$ is a non-decreasing function of t for each fixed x. Thus, an eruption of any magnitude x is more likely to have been recorded if it occurred recently rather than in the distant past;

3. $p(t, x)$ is a non-decreasing function of x for each fixed t. This means that at any point in time, events of a larger magnitude were less likely to be missed than those of smaller magnitude.

This still leaves many possibilities. For this article we have adopted

$$p(t, x) = \left(1 - \frac{v}{x^w}\right) + \frac{v}{x^w} t^b, \tag{7}$$

where the parameters (v, w, b) satisfy $b \geq 0$, $w \geq 0$ and $v < u^w$. Each of the parameters in the model has its own interpretation: v determines the extent to which events are historically censored ($v = 0$ would imply no historical censoring); w determines the extent to which under-reporting is different at different levels ($w = 0$ would imply a constant under-reporting at all levels); b determines the rate of change in under-reporting at different time-points ($b = 1$ would imply a linear change, for example). The overall result is a 6-parameter model, 3 of whose parameters correspond to the extreme value properties of the genuine process of volcanic eruption that has only been partially observed, and three of which correspond to the recording mechanism.

TABLE 1. Maximum likelihood estimates and standard errors of censored point process model applied to volcano catalogue.

	μ	σ	ξ	v	w	b
Estimate	3.289	1.124	−0.239	1.691	0.413	6.971
Standard Error	0.183	0.158	0.047	0.552	0.219	1.25

Maximum likelihood estimates and their standard errors for this model are given in Table 1. Each of v, w and b is significantly different from 0, v and b overwhelmingly so. The results for v and w are especially important, since they confirm the existence of an historical under-reporting ($v \neq 0$), and that the extent of this is greater for events of low magnitude ($w \neq 0$). These conclusions are supported further by a comparison of the maximized log-likelihoods in Table 2.

TABLE 2. Maximized log-likelihood values for different point process sub-models.

	Unconstrained	$v = 0$	$w = 0$
log-lik	−820.96	−890.81	−823.39
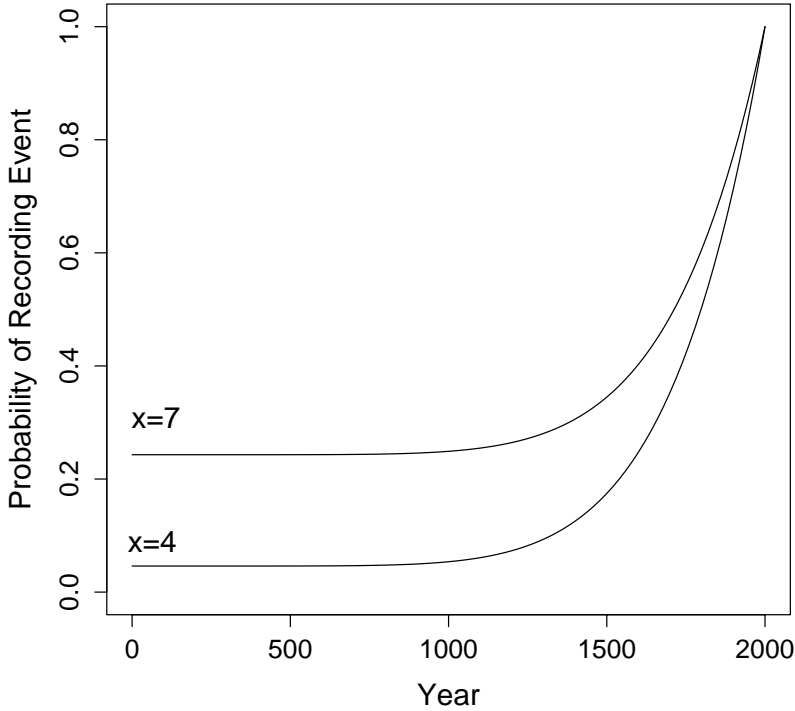
FIGURE 4. Censoring function $p(t, x)$ for each of $x = 4$ and $x = 7$.

The estimated function $p(x, t)$ is plotted in Fig. 4 for two different values of the magnitude x. These suggest a near-constant recording probability at each threshold for the first 1500 years or so, followed by a rapid rise in the rate. This accords with what one might expect: a sharp rise due to the expansion in sociological, scientific and technical facilities that have taken place over the last 500 years. The estimated recording probabilities at the respective magnitudes $x = 4$ and $x = 7$ in the year 0 are around 5% and 25%, emphasising the strength of the estimated under-reporting effect.

The parameterization of our model implies that the parameters (μ, σ, ξ) correspond to the current process of volcanic activity, which is assumed to be recorded perfectly. The corresponding return level curve is plotted in Fig. 5, together with the estimate that would be obtained for the same period of data but ignoring the under-recording mechanism. Though they
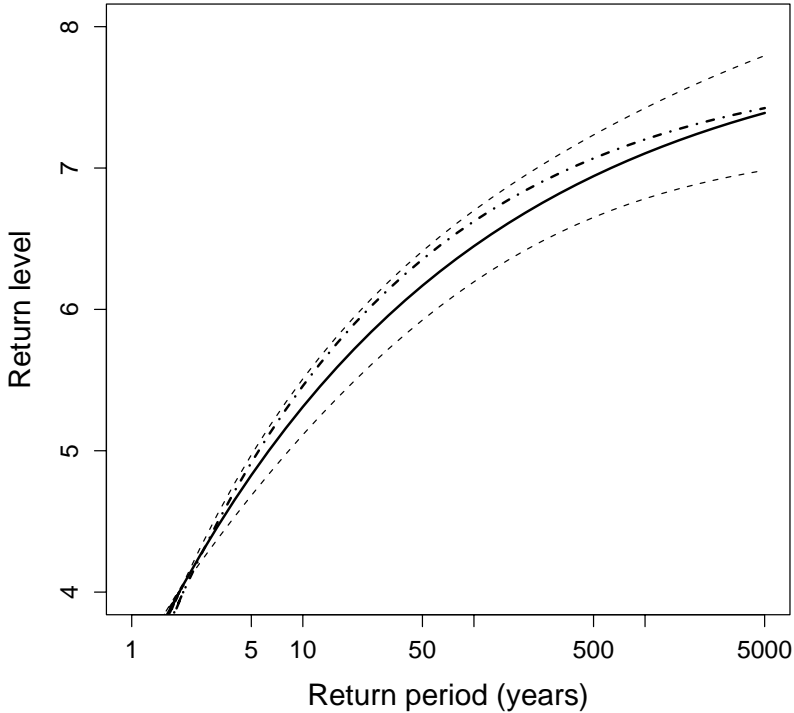
FIGURE 5. Return level curve of volcanic eruption magnitudes. Solid line corresponds to censored point process model, with limits of pointwise 95% confidence intervals shown as dashed curves. Broken-dashed curve corresponds to mis-specified homogeneous Poisson process model.

are significantly different, the differences are not so great in absolute terms, suggesting that the dependence of the censoring term $p(t, x)$ on x is not so severe as to induce much bias if it is ignored in the model estimation. However, this conclusion may be driven in part by the particular choice of parametric model adopted.

One important calculation that can be made on the basis of the fitted model is an estimate of the maximum magnitude achievable in a volcanic eruption. Within the Poisson process model this limit is $X_{\max} = \mu - \sigma/\xi$. Based on the fitted model the maximum likelihood estimate is $\hat{X}_{\max} = 7.99$ with a 95% confidence interval of $[7.1, 8.88]$ obtained via the delta method. Given that the value of 7.1 has already been exceeded, there are certainly better

ways to obtain such an interval. Moreover, given the necessity to formulate questions of volcanic activity within a risk assessment framework, it would be arguably better to do the entire inference in a Bayesian setting. This is one aspect of Claudia Furlan's contribution to these same proceedings.

4   Conclusions

The Poisson process representation for extremes seems a natural framework for modelling the extremes of processes that are subject to some secondary perturbation, in this case, the historical under-recording of events of low magnitude. The model enables all of the available data to be exploited, but avoids the bias that would occur if the under-reporting of weak events were ignored. Our results point to a volcanic activity rate – in the sense of exceeding a level of 3.7 – of roughly once every two years. The maximum feasible eruption size is estimated at around $X_{\max} = 8$, or $X_{\max} = 9$ after taking sampling effects into account. These results are broadly consistent with the volcanological literature (Mason *et al.*, 2004, for example) based both on other statistical analyses, and a geological calculation of the physical limits to volcanic magnitude.

There remain other issues to explore. The choice of parametric model for $p(\cdot, \cdot)$ is essentially arbitrary, and there are probably better ways to handle this aspect. Indeed, given the risk assessment nature of the problem, it may be much better to formulate the whole problem within a Bayesian framework and adopt alternative approaches to the specification of $p(\cdot, \cdot)$. This issue is considered in Claudia Furlan's contribution to these proceedings, together with the various advantages that accrue from a Bayesian approach to the same problem. Other issues that we have not yet looked at include the possibility of modelling individual or groups of volcanoes separately, rather than assuming, as here, that they all have identical stochastic properties, and the possibility of time-dependence in the eruption process, which would violate the Poisson assumptions that we have made here.

In summary, although there are undoubtedly better ways of building the intensity model in (6), the general approach of integrating process knowledge within a statistical model – albeit in a naive way – appears to have produced useful results. Hopefully, this also is a metaphor for the further integration of contemporary statistical thinking into volcanological science.

References

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. London: Springer.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds (with discussion). *Journal of the Royal Statistical Society, B*, **52**, 393-442.

Mason, B. G., Pyle, D. M. and Oppenheimer, C. (2004). The size and frequency of the largest explosive eruptions on Earth. Bulletin of Volcanology. To appear.

Pickands, J. (1971). The two-dimensional Poisson process and extremal processes. *Journal of Applied Probability*, **8**, 745-756.

Simkin, T. and Siebert, L. (1994). *Volcanoes of the World*. Tucson: Geoscience Press.

Smith, R. L. (1989). Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone (with discussion). *Statistical Science*, **4**, 367-393.

Marginal Models: recent developments and applications to categorical time series analysis.

Roberto Colombi[1]

[1] Università di Bergamo - Viale Marconi - 24044 Dalmine Italy; colombi@unibg.it

Abstract: Recently general definitions of marginal interactions and marginal models have been introduced by Bergsma, Rudas (2002), Colombi, Forcina (2001) and by Bartolucci, Colombi, Forcina (2004) that considerably improved the flexibility and interpretability of standard hierarchical log-linear models by allowing interactions to be contrasts of four types of Logits defined within different marginal distributions. This paper reviews these recent contributions and shows their relevance in the context of categorical time series analysis.

Keywords: marginal models, categorical time series, non-normal state space models

1   Introduction

In section two of this paper we review the definition of generalized marginal interactions introduced by Bartolucci, Colombi, Forcina (2004) and we show how these interactions are used to build a class of models which generalizes the Hierarchical Marginal Models previously introduced by Bergsma, Rudas (2002). In section three of this paper the proposed marginal models are used to specify a class of dynamic models for multi-categorical time series and in section four some examples are given. The aim of the work is to show that marginal parameterizations can be easily adapted to the context of categorical time series modelling.

2   Marginal interaction parameters and marginal models

Consider the joint probability function of q response variables $A_1, \ldots, A_q$, with A_j taking values x_j in $\{1, 2, \ldots, a_j\}$. The set of response variables that defines a given marginal distribution will be denoted by the set $\mathcal{M}$ of indices of the corresponding variables and $\mathcal{Q} = \{1, \ldots, q\}$ will refer to the joint distribution. The vector of the $\prod_1^q a_j$ joint probabilities will be denoted by $\boldsymbol{\pi}$.

2.1 Generalized Marginal Interactions

We now introduce the Bartolucci, Colombi, Forcina (2004) definition of interaction parameters which includes the four well known types of logits: *local* (l), *global* (g), *continuation* (c) and *reverse continuation* (r) and the sixteen types of log-odds ratios discussed by Douglas *et al.* (1990). Note that it makes sense to use logits of type *local* both with ordinal and non-ordinal variables but that logits of type *global* and *continuation* can be used only with ordinal variables.

For any *category* $x_j < a_j$, define the event $\mathcal{B}(x_j, 0)$ to be equal to $\{x_j\}$ if the logit is of type local or continuation and to $\{1, \ldots, x_j\}$ for global or reverse continuation logits; similarly, the event $\mathcal{B}(x_j, 1)$ is equal to $\{x_j + 1\}$ if the logit is of type local or reverse continuation and to $\{x_j + 1, \ldots, a_j\}$ for global or continuation logits. Finally define the marginal probabilities:

$$p_{\mathcal{M}}(\boldsymbol{x}_{\mathcal{M}}; \boldsymbol{h}_{\mathcal{M}}) = p(A_j \in \mathcal{B}(x_j, h_j), \ \forall j \in \mathcal{M}),$$

where $\boldsymbol{x}_{\mathcal{M}}$ is a row vector of categories x_j, $j \in \mathcal{M}$, and $\boldsymbol{h}_{\mathcal{M}}$ is a row vector whose elements, h_j, $j \in \mathcal{M}$, are equal to zero or to one. These marginal probabilities are probabilities of a table where the variables $A_j, \forall j \in \mathcal{M}$, have been dichotomized according to the categories: $\mathcal{B}(x_j, 0)$, $\mathcal{B}(x_j, 1)$. The marginal generalized interactions are log-linear contrasts of the previous probabilities and are so defined:

$$\eta_{\mathcal{H};\mathcal{M}}(\boldsymbol{x}_{\mathcal{H}} \mid \boldsymbol{x}_{\mathcal{M}\setminus\mathcal{H}}; \boldsymbol{h}_{\mathcal{M}\setminus\mathcal{H}}) = \sum_{\mathcal{K}\subseteq\mathcal{H}} (-1)^{|\mathcal{H}\setminus\mathcal{K}|} \log p_{\mathcal{M}}(\boldsymbol{x}_{\mathcal{M}}; \boldsymbol{h}_{\mathcal{M}\setminus\mathcal{H}}, \boldsymbol{0}_{\mathcal{H}\setminus\mathcal{K}}, \boldsymbol{1}_{\mathcal{K}}).$$
$$(1)$$

Note that any interaction is defined by the *interaction set* $\mathcal{H}$ of the variables involved, by the marginal distribution $\mathcal{M}$ where it is defined and by the logit type assigned to each variable of $\mathcal{M}$. According to this definition the kind of dichotomy implied by the type of logit adopted for each variable should carry over when defining higher order interactions within the same marginal distribution. As an example consider the bivariate case, $q = 2$, where the continuation logit type is assigned to each variable and the marginals of interest are: $\mathcal{M}_1 = \{1\}$, $\mathcal{M}_2 = \{2\}$ and $\mathcal{M}_3 = \{1, 2\}$. Let π_{ij}, $\pi_{i\cdot}$ and $\pi_{\cdot j}$ denote the joint and marginal probabilities, then

$$\boldsymbol{\eta}_{\{1\};\{1\}}(i) = \ln \frac{p(A_1 \in \mathcal{B}(i, 1))}{p(A_1 \in \mathcal{B}(i, 0))} = \ln \frac{\sum_{n=i+1}^{a_1} \pi_{n\cdot}}{\pi_{i\cdot}},$$

$$\boldsymbol{\eta}_{\{2\};\{2\}}(j) = \ln \frac{p(A_2 \in \mathcal{B}(j, 1))}{p(A_2 \in \mathcal{B}(j, 0)} = \ln \frac{\sum_{n=j+1}^{a_2} \pi_{\cdot n}}{\pi_{\cdot j}},$$

and

$$\boldsymbol{\eta}_{\{1,2\};\{1,2\}}(ij) =$$

$$= \ln \frac{p(A_1 \in \mathcal{B}(i,1), A_2 \in \mathcal{B}(j,1))}{p(A_1 \in \mathcal{B}(i,0), A_2 \in \mathcal{B}(j,1))} - \ln \frac{p(A_1 \in \mathcal{B}(i,1), A_2 \in \mathcal{B}(j,0))}{p(A_1 \in \mathcal{B}(i,0), A_2 \in \mathcal{B}(j,0)} =$$

$$= \ln \frac{\sum_{m=j+1}^{a_2} \sum_{n=i+1}^{a_1} \pi_{nm}}{\sum_{m=j+1}^{a_2} \pi_{im}} - \ln \frac{\sum_{n=i+1}^{a_1} \pi_{nj}}{\pi_{ij}},$$

are continuation log-odds ratios.

2.2   Complete and Hierarchical Families of Interaction Sets

We now examine the problem of allocating the *interaction sets* among the marginals within which they may be defined.

Denote by $\mathcal{F}_m$ the family of interaction sets defined within the marginal distribution $\mathcal{M}_m$. Let also $\mathcal{P}(\mathcal{J})$ be the family of all non empty subsets of $\mathcal{J}$ and $\mathcal{P}_m$ be a short-hand notation for $\mathcal{P}(\mathcal{M}_m)$.

Given a non-decreasing sequence of marginals $\mathcal{M}_1, \ldots, \mathcal{M}_s$, a family of interactions sets is called complete and hierarchical if (i) any interaction set is defined in one marginal distribution $\mathcal{M}_m$, (ii) $\mathcal{F}_1 = \mathcal{P}_1$ and $\mathcal{F}_m = \mathcal{P}_m \setminus \bigcup_{h<m} \mathcal{F}_h$.

The previous definition implies that $\mathcal{M}_s = \mathcal{Q}$, that $\mathcal{M}_m \in \mathcal{F}_m$, for every m, that every family $\mathcal{F}_m$ is a non-empty ascending class of subsets of $\mathcal{M}_m$ and that every interaction is defined within only one marginal distribution. In the following, for every interaction set $\mathcal{I} \in \mathcal{M}_m$ of a complete hierarchical family of interactions sets, we will consider only the interactions:

$$\eta_{\mathcal{I}; \mathcal{M}_m}(\boldsymbol{x}_{\mathcal{I}}) = \eta_{\mathcal{I}; \mathcal{M}_m}(\boldsymbol{x}_{\mathcal{I}} \mid \boldsymbol{1}_{\mathcal{M}_m \setminus \mathcal{I}}; \boldsymbol{0}_{\mathcal{M}_m \setminus \mathcal{I}})$$

where the conditioning variables of $\mathcal{M}_m \setminus \mathcal{I}$ are fixed to their first category. When all the conditioning variables in $\mathcal{M}_m \setminus \mathcal{I}$ have assigned logits of type local Bartolucci, Colombi, Forcina (2004) showed that the interactions $\eta_{\mathcal{I}; \mathcal{M}_m}(\boldsymbol{x}_{\mathcal{I}} \mid \boldsymbol{x}_{M_m \setminus \mathcal{I}}; \boldsymbol{h}_{M_m \setminus \mathcal{I}})$ are linear functions of the interactions $\eta_{\mathcal{H}; \mathcal{M}_m}(\boldsymbol{x}_{\mathcal{H}})$, $\mathcal{H} \supseteq \mathcal{I}$, so that at least in this case there is no restriction in limiting the attention to these parameters.

2.3   Complete and Hierarchical Marginal Parametrizations

The interactions $\eta_{\mathcal{I}; \mathcal{M}_m}(\boldsymbol{x}_{\mathcal{I}})$ associated to a complete hierarchical family of interactions may be arranged into the vector $\boldsymbol{\eta}$ which may be explicitly written in matrix form as

$$\boldsymbol{\eta} = \boldsymbol{C} \log(\boldsymbol{M} \boldsymbol{\pi}), \tag{2}$$

where the rows of $\boldsymbol{C}$ are contrasts and $\boldsymbol{M}$ is a matrix of zeros and ones which sums the probabilities of appropriate cells to obtain the necessary marginal probabilities of the type described by (2.1). A detailed description of these matrices is given by Colombi, Forcina (2001). Bartolucci, Colombi, Forcina (2004) showed that (2) is invertible. The result extends

the Bergsma, Rudas (2002) important contribution on marginal models and earlier works of Lang, Agresti (1994), Glonek, McCullagh (1995) and Glonek (1996). Parameters defined by a function of the joint probabilities of the type (2) have a long history starting from the seminal works of Grizzle *et al.* (1969) and of Forthofer, Koch (1973) and here we stress the fact that the representation of the link function (2) is important, in the context of maximum likelihood estimation, both from the theoretical point of view and from the computational point of view. The importance of the representation will carry over also to the context of categorical time series as it will be shown in the next section.

A parameterization of the joint probabilities in term of the generalized marginal interactions $\eta_{\mathcal{I};\mathcal{M}_m}(\boldsymbol{x}_{\mathcal{I}})$ defined as above will be called *complete hierarchical marginal parameterization.*

The advantages of a marginal parameterization with respect to the loglinear one come from the flexibility in the choice of the interactions and from the interpretability of the parameters. Marginal parametrizations allow a direct and straightforward parameterization of the marginal probabilities of interest and in the framework of a marginal parameterization it is easier to state that a given marginal distribution is stochastically larger than another or that the strength of the dependence between two variables increase with a third variable or that two variables are marginally independent or positively associated. In fact these hypotheses can be defined by linear inequality and equality constraints on generalized marginal interactions as shown in Dardanoni, Forcina (1998), Bartolucci, Forcina, Dardanoni (2001), Colombi, Forcina (2001) and Bartolucci, Colombi, Forcina (2004). Moreover complete hierarchical marginal parameterizations are very useful in parametrizing block recursive models as shown by Bartolucci, Colombi, Forcina (2004).

As an example consider the seemeengly unrelated logit regressions represented by the dashed edges graph of figure 5.3(a) of Cox, Wermuth (1996); under this model the variables A_3 and A_4 are explanatory for the variables A_1 and A_2, A_2 is independent from A_3 given A_4 and A_1 is independent from A_4 given A_3. The model can be parametrized choosing the complete hierarchical parameterization defined by the marginals $\mathcal{M}_1 = \{3,4\}$, $\mathcal{M}_2 = \{1,3,4\}$, $\mathcal{M}_3 = \{2,3,4\}$, $\mathcal{M}_4 = \{1,2,3,4\}$, and the constraints:

$$\boldsymbol{\eta}_{\{2,3\};\{2,3,4\}}(\boldsymbol{i}_{\{2,3\}}) = 0, \quad \boldsymbol{\eta}_{\{2,3,4\};\{2,3,4\}}(\boldsymbol{i}_{\{2,3,4\}}) = 0,$$

$$\boldsymbol{\eta}_{\{1,4\};\{1,3,4\}}(\boldsymbol{i}_{\{1,4\}}) = 0, \quad \boldsymbol{\eta}_{\{1,3,4\};\{1,3,4\}}(\boldsymbol{i}_{\{1,3,4\}}) = 0.$$

If the four categorical variables are ordinal it is sensible to choose logits of type global for A_3 and A_4 within $\mathcal{M}_1$ and for A_1 and A_2 within $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$. As explained in Bartolucci, Colombi, Forcina (2004), who gave a general description of block recursive models of this type, it is convenient to use logits of type local for the explanatory variables A_3 and A_4 within $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$.

Furthermore together with the previous equality constraints the following inequality constraints:

$$\boldsymbol{\eta}_{\{2,4\};\{2,3,4\}}(\boldsymbol{i}_{\{2,4\}}) \geq 0, \quad \boldsymbol{\eta}_{\{1,3\};\{1,3,4\}}(\boldsymbol{i}_{\{1,3\}}) \geq 0,$$

state that the distributions of A_2 conditioned by the explanatory variables are stochastically increasing with the categories of A_4 and that the conditional distributions of A_1 are stochastically increasing with the categories of A_3. The problem of testing linear inequality constraints on marginal parameters has been discussed by Dardanoni, Forcina (1998), Colombi, Forcina (2001) and by Bartolucci, Colombi, Forcina (2004).

3   Multinomial State Space Models

In this section marginal models are used to introduce a class of dynamic models for multicategorical time series. For a survey of the state of art on categorical time series analysis see Fahrmeir, Tutz (1994), MacDonald, Zucchini (1997), Davis, Wang (1999) and Kedem, Fokianos (2002). Let $\boldsymbol{\pi}_t$ be the vector of the joint probabilities of the categories of q categorical variables given the information set $\mathcal{F}_{t-1}$ available at time t. We parametrize the joint probabilities $\boldsymbol{\pi}_t$ by inverting at time t the link function:

$$\boldsymbol{\eta}_t = \boldsymbol{C} \ln \boldsymbol{M} \boldsymbol{\pi}_t, \tag{3}$$

where the vector of marginal parameters is a linear function of time varying regressors: $\boldsymbol{\eta}_t = \boldsymbol{X}_t \boldsymbol{\beta}_t$ and where $\boldsymbol{\beta}_t$ changes according to a standard normal transition model:

$$\boldsymbol{\beta}_t = \boldsymbol{F} \boldsymbol{\beta}_{t-1} + \boldsymbol{H} \boldsymbol{\varepsilon}_t. \tag{4}$$

Here $\boldsymbol{\varepsilon}_t$ are independent multinormal random variables with null expected value and unknown diagonal variance matrix $\boldsymbol{Q}$. For a discussion of state space models for categorical data and count data see Kedem, Fokianos (2002), Durbin, Koopmann (1997) and Fahrmeir, Tutz (1996). Special cases of the previous general model (for example $\boldsymbol{X}_t = \boldsymbol{I}$, $\boldsymbol{H} = \boldsymbol{I}$ and $\boldsymbol{F} = \boldsymbol{I}$) are easily obtained and the advantage of defining the transition model in function of the marginal parameters rather than the log-linear ones come from the fact that the normal transition model applied to log-linear parameters is often difficult to interpret. On the contrary the transition model applied to marginal interactions and in first place to marginal Logits is very easy to interpret and a more natural and direct modelling strategy. Moreover in the context of categorical time series many important non-Granger causality type hypotheses, which state that a set of categorical variables doesn't depend on the past of another set of variables, given $\mathcal{F}_t$, are equivalent to linear hypotheses on marginal interactions and this fact enhances the importance of marginal models in this context. Finally in the

context of marginal models it is easier to distinguish between hypotheses of simultaneous independence between categorical variables and hypotheses of independence of a categorical variable from the past of the others. These advantages of marginal modelling have been firstly pinpointed by Giordano (2003) in the context of models for the joint transition probabilities of multivariate Markov Chains and the problem of testing Granger non-causality under Markov assumptions was firstly considered by Bouissou *et al* (1986). The important topic of modelling multivariate Markov Chain was started by the works of Fahrmeir, Kaufmann (1987) and Kaufmann (1987) and generalized to a less stringent assumption than the one of Markovianity by Fokianos, Kedem (1998). Hidden Markov models (MacDonald, Zucchini, 1997) can also be considered in this context by substituting the normal transition model (4) with the following one:

$$\boldsymbol{\beta}_t = S_t \boldsymbol{\delta}_1 + (1 - S_t) \boldsymbol{\delta}_2$$

where the binary variable S_t indicates the state at time t of a two state markov Chain.

In this last case the maximum likelihood estimates are easily computed (MacDonald, Zucchini 1997, Krolzig 1997) and in the case of a normal transition model maximum likelihood estimation of the unknown parameters of the multivariate normal distribution of $\boldsymbol{\varepsilon}_t$ can be performed by the Montecarlo likelihood method of Durbin, Koopman (1997, 2001) or by the Montecarlo EM algorithm of Chan, Ledolter (1995). Less computationally demanding methods are the EM-type algorithm of Fahrmeir, Wagenpfeil (1997) and the method based on the maximization of an approximation of the log-likelihood of Durbin, Koopman (1997, 2001). Note that in the case of marginal models all the previous methods are more computationally demanding, than in the cases previously considered, because at every iteration the relation $\boldsymbol{\eta}_t = \boldsymbol{C} \ln \boldsymbol{M} \boldsymbol{\pi_t}$ must be inverted for every t.

The asymptotic properties of the M.L. estimator of the unknown parameters in the case of a latent Markov Chain with time homogeneous transition probabilities follow from the results of Bickel, Ritov, Ryden (1998) on Hidden Markov Models. The asymptotic normality of the M.L. estimators for non-normal state-space model is discussed in Jensen, Petersen (1999).

3.1   Bivariate Markov Driven Marginal Models

Often multi-categorical time series exhibit two different regimes. The starting time and the length of the spells in the regimes are random. To model the different behavior of the time series under the two regimes the parameters of a Marginal Model can be let to depend on the state of an unobservable Markov Chain which models the transitions between the regimes. A latent variable problem arises because the regime is not an observable variable. More precisely the model must consist of two parts:

I) a *Marginal Model* which specifies the joint probabilities of the categories of the variables at time t given the categories of the variables at the previous *lag* times $t-1, t-2, ..., t-lag$, given the values (at time $t-1$) of a vector of regressors $\mathbf{x}_{t-1}$ and given the regime S_t at time t ($S_t = 1$ or $S_t = 0$ in the case of two regimes).

II) a *two states Markov Chain* that models the history of the unobservable regimes S_t.

According to this model the observed multi-categorical time series is not Markovian, however conditionally on the series $\{S_t\}$ of the regimes it is a Markov Chain of order *lag*.

Here we examine the case of a bivariate categorical time series $\{A_{1,t}, A_{2,t}\}$. The joint probability function of $A_{1,t}$ and $A_{2,t}$ conditionally on the past can be specified by a log-linear model. Let $\mathbf{Z}_t$ be the vector of predetermined variables at time t and of the unobservable regime S_t. Then, the log-linear model:

$$\ln \pi_{ij,t} = \lambda_t + \lambda_{i,t}^{A_1} + \lambda_{j,t}^{A_2} + \lambda_{ij,t}^{A_1 A_2},$$
$$i = 1, 2, ..., a_1, j = 1, 2, ..., a_2,$$

could be introduced by allowing the interaction parameters *lambda* to depend on the vector $\mathbf{Z}_t$ of predetermined variables. This approach doesn't allow a direct parameterization of the marginal probabilities $\pi_{i.,t}$, $\pi_{.j,t}$. For this reason we prefer to parametrize the marginal probabilities directly with univariate logit Models. For example the Continuation logit Parameterization (Colombi, Forcina 1999) for the marginal probabilities is given by the following formulae:

$$\pi_{i.,t} = \frac{\exp\{-\eta_{1,t}(i)\}}{\prod_{m=1}^{i} [1 + \exp\{-\eta_{1,t}(m)\}]}, i = 1, 2, ..., a_1 - 1,$$

$$\pi_{.j,t} = \frac{\exp\{-\eta_{2,t}(j)\}}{\prod_{m=1}^{j} [1 + \exp\{-\eta_{2,t}(m)\}]}, j = 1, 2, ..., a_2 - 1.$$

Here we have slightly simplified the notation of interactions given in section two by omitting curly brackets and the indication of the marginal within which the interaction is defined. The Continuation Logits $\eta_{1,t}(i)$ and $\eta_{2,t}(j)$ depend on the vector of predetermined variables $\mathbf{Z}_t$ according to linear predictors of the type commonly used in the context of logit regression (see section 4 for an example). Note that the Continuation logit of a categorical variable may depend also on the past of the other categorical variable. The joint probabilities $\pi_{ij,t}$ are specified by the marginal continuation logits and by the logarithms of the Continuation Odds Ratios (Colombi, Forcina 1999):

$$\eta_{12,t}(ij) = \ln \frac{\pi_{ij,t} \cdot \sum_{m=i+1}^{a_1} \sum_{n=j+1}^{a_2} \pi_{mn,t}}{\sum_{m=i+1}^{a_1} \pi_{mj,t} \cdot \sum_{n=j+1}^{a_2} \pi_{in,t}},$$
$$i = 1, 2, ..., a_1 - 1, \quad j = 1, 2, ..., a_2 - 1.$$

The following hypotheses on the Continuation Odds Ratios are relevant:

$$\begin{aligned}
\eta_{12,t}(ij) &= \eta_{12}(ij), \\
\eta_{12,t}(ij) &= \eta_{12}(ij) + \rho S_t.
\end{aligned}$$

Both are hypotheses of constant association in the sense that the Continuation Odds Ratios do not depend on the past of A_1 and A_2 and on a vector of regressors $\mathbf{x}_{t-1}$. In the first case, the Odds Ratios are also regime independent, whereas in the second case the Continuation Odds Ratios depend on the latent regime but the effect of the regime is the same for all i and j $(i = 1, 2, ..., a_1 - 1; j = 1, 2, ..., a_2 - 1)$. A more parsimonious model is given by the following hypotheses of Uniform Constant association:

$$\begin{aligned}
\eta_{12,t}(ij) &= \eta_{12}, \\
\eta_{12,t}(ij) &= \eta_{12} + \rho S_t.
\end{aligned}$$

Finally the transition probabilities of the Hidden Markov Chain $p_{00t} = p(S_{t+1} = 0|S_t = 0)$ and $p_{11t} = p(S_{t+1} = 1|S_t = 1)$ can assumed to be function of a vector of regressors $\mathbf{x}_{t-1}$ according to the logit Models:

$$\ln \frac{p_{iit}}{1 - p_{iit}} = \alpha_{0i} + \boldsymbol{\alpha}'_{1i}\mathbf{x}_{t-1}, i = 0, 1. \tag{5}$$

The case of a time homogeneous transition matrix is obtained by putting $\boldsymbol{\alpha}_{1i} = \mathbf{0}$, $i = 0, 1$.

Given the marginal continuation logits and the Continuation Odds Ratios the joint probabilities $\pi_{ij,t}$ can be computed with the iterative algorithm introduced by Colombi, Forcina (1999) and described in Colombi, Zanarotti (2002).

Let $\boldsymbol{\vartheta}' = [\alpha_{00}, \boldsymbol{\alpha}_{10}, \alpha_{01}, \boldsymbol{\alpha}_{11}, \boldsymbol{\theta}']$ be the vector of the parameters to be estimated where $\boldsymbol{\theta}$ is the vector of the parameters of the bivariate marginal model. Given the parameters, the BLHK filter and smoother (Krolzig, 1997) can be used to marginalize with respect the unobservable Markov Chain and to compute the log-likelihood at every iteration of the Fisher Scoring algorithm.

3.2 State Space Trend Models for categorical data

Marginal State Space Models for categorical data can be specified in many ways thanks to the flexibility of the definition of $\boldsymbol{\eta}_t$ and of the transition model: $\boldsymbol{\eta}_t = \boldsymbol{X}_t\boldsymbol{\beta}_t$, $\boldsymbol{\beta}_t = \boldsymbol{F}\boldsymbol{\beta}_{t-1} + \boldsymbol{H}\boldsymbol{\varepsilon}_t$. A first important and useful case is given by the $(k-1)$- polynomial stochastic trend where some components $\eta_{i,t}$ change according to the transition model:

$$\boldsymbol{\beta}_{i,t} = \boldsymbol{F}\boldsymbol{\beta}_{i,t-1} + \boldsymbol{\varepsilon}_t \quad \eta_{i,t} = \beta_{1,it}$$

and $\boldsymbol{F}$ is a $k \cdot k$ upper triangular matrix of ones. A second important example is the case of k order random walk where some components $\eta_{i,t}$ of $\boldsymbol{\eta}_t$ change according to the transition model:

$$\boldsymbol{\beta}_{i,t} = \boldsymbol{F}\boldsymbol{\beta}_{i,t-1} + \boldsymbol{h}\boldsymbol{\varepsilon}_{i,t}$$
$$\eta_{i,t} = \beta_{1,it}.$$

here $\boldsymbol{h}$ is the first column of a $k \cdot k$ identity matrix and $\boldsymbol{F}$ is a $k \cdot k$ identity matrix with the first row replaced by the row vector $\boldsymbol{c}$, the $i - th$ element of which is $c_i = (-1)^{i-1} \begin{pmatrix} k \\ i \end{pmatrix}$. These models are useful to model local-trends for logits defined within different marginals. For example in the Bivariate Case introduced in the previous section a local level model (k=1) can be applied to the two marginal continuation logits:

$$\eta_{1,t}(i) = \eta_{1,t-1}(i) + \varepsilon_{1,t},$$
$$\eta_{2,t}(i) = \eta_{2,t-1}(i) + \varepsilon_{2,t}.$$

4   Ground O3 and CO data analysis

The Hidden Markov models described in section 3.1 are used to analyze daily levels of ground $O3$ (variable A_{1t}) and CO concentration (variable A_{2t}) both with three categories (*low (1), normal(2)* and *high(3))*. Data are taken by San Giorgio (Bergamo-Italy) measurement unit from 1997 to 1999. In this application the covariates that affect the continuation Logits are: *temperature* and *solar radiation*.
The general effects of the linear predictors are assumed to change according to the hidden regime and the other parameters (additive effects, interactions, regression coefficients) are regime independent. More precisely the most general linear predictor used for the $\eta_{1,t}(i), i = 1, 2, ..., a_1 - 1$ is:

$$\eta_{1,t}(i) = \left(\mu_j^{(0)} + \delta_j S_t \right) +$$
$$+ \left(\sum_{l=1}^{lag} \sum_{m=1}^{2} \theta_{ml}^{A_1} I_{\{A_{1,t-l}=m\}} + \sum_{l=1}^{lag} \sum_{m=1}^{2} \theta_{ml}^{A_2} I_{\{A_{2,t-l}=m\}} \right) +$$
$$+ \left(\sum_{l=2}^{lag} \sum_{m=1}^{2} \delta_{ml}^{A_1} \prod_{k=t-l}^{t-1} I_{\{A_{1,k}=m\}} + \sum_{l=2}^{lag} \sum_{m=1}^{2} \delta_{ml}^{A_2} \prod_{k=t-l}^{t-1} I_{\{A_{2,k}=m\}} \right) +$$
$$+ \beta_1 x_{1t} + \beta_2 x_{2t}.$$

A similar predictor is used for the $\eta_{2,t}(j), j = 1, 2, ..., a_2 - 1$. In the first column of Table 1 it is given the number LAG of past pollutant levels that

TABLE 1. Switching Bivariate Marginal Models (O3 and CO)

lag	link	association	log-lik.	n. par.
1	add.	$\eta_{12,t}(ij) = 0$	-919.08	18
2	add.	$\eta_{12,t}(ij) = 0$	-894.08	26
3	add.	$\eta_{12,t}(ij) = 0$	-873.49	34
4	add.	$\eta_{12,t}(ij) = 0$	-858.88	42
4	add.+int.	$\eta_{12,t}(ij) = 0$	-846.27	78
4	add.+int.	$\eta_{12,t}(ij) = \eta_{12}$	-846.26	79
4	add.+int.	$\eta_{12,t}(ij) = \eta_{12}(ij)$	-845.22	82
4	add.+int.+reg.	$\eta_{12,t}(ij) = \eta_{12}(ij)$	-842.52	86

TABLE 2. One step forecasts-O3

$\frac{predicted\rightarrow}{observed\downarrow}$	low	normal	high.	tot.
low	**704**	61	0	765
normal	86	**189**	3	278
high.	0	12	**5**	17
tot.	790	262	8	1060

TABLE 3. One step forecasts-CO

$\frac{predicted\rightarrow}{observed\downarrow}$	low	normal	high	tot.
low	**152**	102	0	254
normal	56	**709**	5	770
high	0	23	**13**	36
tot.	208	834	18	1060

affects the current one. In the second column the linear predictor used is described (*add.* means that the effect of the LAG previous levels is additive and *add.+int.* means that interactions between time adjacent past levels of the same pollutant are also allowed and *add.+int.+reg.* is the general case where also the effects of the covariates *temperature* and *solar radiation* are introduced). In the third column the type of association between CO and O3, given the past levels and the hidden regime, is described. In the fourth column the value of the log-likelihood is reported and in the last column the number of parameters is given. For all the models considered the transition probabilities of the Hidden Markov Chain are time invariant. In the last two tables the one-step predicted levels are crossed with the actual ones, using the model in the last row of Table 1.

In Table 4 the results obtained by using some State Space Trend Models

TABLE 4. Bivariate State Space Models (O3 and CO)

model	number of states	log-lik.
M_1	8	-117.558
M_2	5	-118.444
M_3	12	-115.358
M_4	9	-116.771

introduced in section 3.2 are reported. In this case only the first 100 observations were used, covariates effects were not included and local logits and local odds-ratios were used instead of the continuation ones. In the case of the first model M_1 the four local logits and the four local odds-ratios that parametrize the joint distribution at time t changes according to a random walk. In model M_2 the four odds ratios are assumed to be equal and the five parameters still changes according to a random walk. According to model M_3 the four odds ratios changes according to a random walk and the four logits changes according to a local level local trend model (local polynomial of order one). In model M_4 the transition equation for the logits is as in model M_3 and the four odds-ratios are equal and change according to a random walk. Initial states have been treated as unknown parameters so that the number of parameters to be estimated is twice the number of states. The method based on the maximization of the approximate log-likelihood of Durbin, Koopman (2001) were used but after convergence the log-likelihood was computed with the importance-sampling method of Durbin, Koopman (2001).

References

Bartolucci, F., Colombi, R., Forcina, A. (2004). An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints, submitted to *Annals of Statistics*.

Bartolucci, F., Forcina, A., Dardanoni, V. (2001). Positive quadrant dependence and marginal modelling in two-way tables with ordered margins, *Journal of the American Statistical Association*, 96, pp. 1497-1505.

Bergsma, W. P., Rudas, T. (2002). Marginal models for categorical data, *Annals of Statistics*, 30, pp. 140-159.

Bickel, P.J., Ritov, Y., Ryden, T. (1998). Asymptotic normality of the max-
    imum likelihood estimator for general hidden Markov models. *The
    Annals of Statistics*. 26, pp. 1614-1635.

Bouissou, B., Laffont, J., Vuong, H. (1986) Tests of noncausality under
    Markov assumptions for qualitative panel data. *Econometrica*, 54,
    pp. 395-414.

Chan, S., Ledolter, J. (1995). Monte Carlo EM Estimation for Time Se-
    ries Models Involving Counts, *Journal of The American Statistical
    Association*, 90, pp. 242-252.

Colombi, R., Forcina, A. (1999). An instance of generalized log-linear mod-
    els with inequality constraints: the continuation logit parameteriza-
    tion, *Proceedings of the 14th International Workshop on Statistical
    Modelling*, edited by H. Friedl., Gratz.

Colombi, R., Forcina, A. (2001). Marginal regression models for the anal-
    ysis of positive association of ordinal response variables, *Biometrika*,
    88, pp. 1007-1019.

Colombi, R., Zanarotti, C. (2002). A markov driven bivariate logit model.
    Studi in onore di Angelo Zanella, edited by Frosini B.V., Magagnoli
    U., Boari G., pp 125-135, Vita e Pensiero, Milano.

Cox, R., Wermuth, N. (1996). *Multivariate dependencies, Models analysis
    and interpretation*, Chapman Hall, London.

Dardanoni, V., Forcina, A. (1998). A Unified approach to likelihood infer-
    ence on stochastic orderings in a nonparametric context, *Journal of
    the American Statistical Association*, 93, pp. 1112-1123.

Davis, R., Wang, Y. (1999). Modelling Time series of Count Data, *Asymp-
    totics, Nonparametrics and Time Series* edited by Ghosh S., Marcel
    Dekker, New York.

Douglas, R., Fienberg, S. E., Lee, M. T., Sampson, A. R., Whitaker, L. R.
    (1990). Positive dependence concepts for ordinal contingency tables,
    in *Topics in statistical dependence*, edited by Block H. W., Sampson
    A. R. and Sanits T. H., Institute of Mathematical Statistics, Lecture
    Notes, Monograph series, Haywar, California.

Durbin J., Koopman, S. J. (1997). Monte Carlo maximum likelihood es-
    timation for non-Gaussian state space models, *Biometrika*, 84, pp.
    669-684.

Durbin, J., Koopman, S. J. (2001). *Time Series Analysis by State Space
    Methods*, Oxford University Press, New York.

Fahrmeir, L., Kaufmann, H. (1987). Regression models for nonstationary categorical time series, *Journal of Time Series Analysis*. 8, pp. 147-160.

Fahrmeir, L., Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models* Springer, Berlin.

Fahrmeir, L., Wagenpfeil, W. (1997). Penalized Likelihood estimation and iterative Kalman smoothing for non-Gaussian dynamic regression models, *Computational Statistics and data analysis*, 24, pp.295-320

Fokianos, K., Kedem, B. (1998). Prediction and classification of non-stationary categorical time series. *Journal of Multivariate Analysis*, 67, pp. 277-296.

Forthofer, R., Koch, G. (1973). An analysis of compounded functions of categorical data, *Biometrics*, 29, pp. 143-157.

Giordano, S. (2003). *Modelli Parametrici per catene di Markov bivariate*, Tesi di Dottorato, XV ciclo, Università di Milano-Bicocca, Milano.

Glonek, G. (1996). A class of regression models for multivariate categorical responses, *Biometrika*, 83, pp. 15-28.

Glonek, G., McCullagh, P. (1995). Multivariate Logistic Models, *Journal of the Royal Statistical Society B*, 57, pp. 533-546.

Grizzle, J., Starmer, F., Koch, G. (1969). Analysis of categorical data by linear models, *Biometrics*, 25, pp. 489-505.

Jensen, J., Petersen, N. (1999). Asymptotic normality of the maximum likelihood estimator in State Space models. *Annals of Statistics*, 27, pp. 514-535.

Kaufmann, H. (1987). Regression models for nonstationary categorical time series: asymptotic estimation theory, *The Annals of Statistics*, 15, pp. 79-98.

Kedem, B., Fokianos, K. (2002). *Regression Models for Time Series Analysis*, Wiley, New York.

Krolzig, M. (1997). *Markov Switching Vector Autoregression*, Springer Berlin.

Lang, J., Agresti, A. (1994). Simultaneously modelling the joint and marginal distributions of multivariate categorical responses, *Journal of the American Statistical Association*, 89, pp. 626-632.

MacDonald, I., Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete valued Time Series*, Chapman Hall, London.

Generalized Linear Latent and Mixed Models with Composite Links and Exploded Likelihoods

Anders Skrondal[1] and Sophia Rabe-Hesketh[2]

[1] Norwegian Institute of Public Health, Oslo (`anders.skrondal@fhi.no`)
[2] University of California, Berkeley

Abstract: Applications of composite links and exploded likelihoods for generalized linear latent and mixed models are explored.

1 Introduction

Instead of linking the expectation of each observation with a single linear predictor as in generalized linear models, it is often useful to link it with a composite function of several linear predictors. Moreover, each likelihood contribution can sometimes be exploded into a product of terms.

We explore how these tools can be used to extend 'Generalized Linear Latent And Mixed Models' or GLLAMMs (Rabe-Hesketh, Skrondal and Pickles, 2004a; Skrondal and Rabe-Hesketh, 2004). Applications considered include discrete time frailty models, item response models for ordinal items, unfolding models for attitudes, small area estimation with census information, measurement models combining discrete and continuous latent variables, ability testing with guessing, sensitivity analysis of the assumption of normal random effects, and zero-inflated Poisson models.

2 Generalized Linear Models

Let y_i be the response and $\mathbf{x}_i$ explanatory variables for unit i, and define the conditional expectation of the response given the covariates as μ_i, i.e. $\mu_i \equiv \mathrm{E}[y_i|\mathbf{x}_i]$. Generalized linear models can be specified as

$$\mu_i = g^{-1}(\nu_i),$$

where $g^{-1}(\cdot)$ is an inverse link function, $\nu_i = \mathbf{x}_i'\boldsymbol{\beta}$ is a linear predictor and $\boldsymbol{\beta}$ are fixed effects. The specification is completed by choosing a conditional distribution for the responses y_i given the conditional expectations μ_i, $f(y_i|\mu_i)$, from the exponential family.

3   Exploded likelihoods and composite links

3.1   Exploded likelihoods

Generalized linear models can be extended to handle *multivariate responses* y_{it}, $t = 1, \ldots, T$, for each unit. The responses may be of mixed types combining different links and families, for instance a Poisson distributed count and a logistically distributed dichotomous response. Dependence can be modelled by including latent variables (random effects and/or factors) in the linear predictors; see Section 4. Given the corresponding vectors of conditional means $\boldsymbol{\mu}_i$ (which depend on the latent variables), the joint conditional distribution of the vector of responses $\mathbf{y}_i$ is

$$\Pr(\mathbf{y}_i | \boldsymbol{\mu}_i) = \prod_{t=1}^{T} f_t(y_{it} | \mu_{it}). \tag{1}$$

We now distinguish between two types of artificial multivariate responses where the response is univariate but individual likelihood contributions are nevertheless 'exploded' into product terms:

Phantom responses A univariate response y_i can in some cases be represented by S phantom responses y_{it} entering the likelihood (1) as if they were truly multivariate responses.

Phantom responses can be used for the Luce-Plackett model for rankings where the likelihood contribution of a ranking is the product of successive multinomial logit choice probabilities among remaining alternatives (e.g. Skrondal and Rabe-Hesketh, 2003). Another example is survival analysis based on data exploded into risk sets, for instance the Cox proportional hazard model implemented via Poisson regression and the complementary log-log model for discrete time hazards (e.g. Skrondal and Rabe-Hesketh, 2004, Ch.2).

Mutually exclusive responses A univariate response y_i can sometimes be represented by one of S mutually exclusive responses y_{it} having distributions $f_t(y_{it} | \mu_{it})$ from generalized linear models. For the case of $T = 2$ the likelihood can be written as

$$\Pr(\mathbf{y}_i | \boldsymbol{\mu}_i) = f_1(y_{i1} | \mu_{i1})^{1-\delta_i} f_2(y_{i2} | \mu_{i2})^{\delta_i},$$

where the indicator δ_i picks out the appropriate component.

A simple example is a log-normal survival model with right-censoring. Let $\mathbf{x}_i' \boldsymbol{\beta}$ be the linear predictor, y_{i1} the log survival time if the event is observed for i ($\delta_i = 0$) and y_{i2} the censoring time if the event is censored ($\delta_i = 1$). The likelihood contribution then becomes either a normal distribution with identity link and linear predictor $\mathbf{x}_i' \boldsymbol{\beta}$, $f_1(y_{i1} | \mu_{i1}) = \phi(y_{i1}; \mu_i, \sigma^2)$, or a Bernoulli distribution with a (scaled) probit link and linear predictor $\mathbf{x}_i' \boldsymbol{\beta}$, $f_2(y_{i2} | \mu_{i2}) = \Phi(\frac{\mathbf{x}_i' \boldsymbol{\beta} - y_{i2}}{\sigma})$. Here, $\Phi(\cdot)$ is the cumulative standard normal distribution and $-y_{i2}$ is treated as an offset.

3.2   Composite links

Thompson and Baker (1981) suggested linking the expectation μ_i with a composite function of several linear predictors instead of a function of a single linear predictor as in generalized linear models.

Simple composite links In this case the expectation μ_i is a weighted sum of inverse links with known weights w_{ir},

$$\mu_i \;=\; \sum_r w_{ir}\, g_r^{-1}(\nu_{ir}),$$

where ν_{ir} is the rth linear predictor for unit i and $g_r^{-1}(\cdot)$ an inverse link function.

A simple example of composite links are cumulative models for categorical responses with S ordered response categories $s = 1, \ldots, S$, which can be expressed as

$$\Pr(y_i > s | \mathbf{x}_i) \;=\; g^{-1}(\nu_i - \kappa_s), \quad s = 1, \ldots, S-1$$

where κ_s are threshold parameters and the inverse link function is a cumulative distribution function such as the standard normal, logistic or extreme value distributions. The response probabilities can be written as a composite link,

$$\Pr(y_i = s | \mathbf{x}_i) \;=\; g^{-1}(\nu_{i,s-1}) - g^{-1}(\nu_{is}), \quad \nu_{is} = \nu_i - \kappa_s, \quad s = 1, \ldots, S, \quad (2)$$

where $\kappa_0 = -\infty$ and $\kappa_S = \infty$ so that $g^{-1}(\nu_{i0}) = 1$ and $g^{-1}(\nu_{iS}) = 0$. An advantage of the composite link formulation is that left and right-censoring, or even interval censoring of an ordinal response are easily accommodated. This is particularly useful for discrete time survival data.

Bilinear composite links A first extension is to consider unknown linear functions of inverse links, replacing the known constants w_{ir} with products of the constants and unknown parameters α_r, giving

$$\mu_i \;=\; \sum_r \alpha_r w_{ir}\, g_r^{-1}(\nu_{ir}).$$

A second extension is to let the expectation be some (not necessarily linear) function $h\{\cdot\}$ of the above sum,

$$\mu_i \;=\; h\{\sum_r \alpha_r w_{ir}\, g_r^{-1}(\nu_{ir})\}.$$

General composite links In this case general functions $f_{ir}[g_r^{-1}(\nu_{ir})]$ replace $w_{ir}\, g_r^{-1}(\nu_{ir})$ in the above expressions.

4 Generalized Linear Latent and Mixed Models

4.1 Generalized Linear Mixed Models (GLMMs)

A crucial assumption of generalized linear models is that the responses of different units i are independent given the covariates $\mathbf{x}_i$. This assumption is often unrealistic since data are frequently of a multilevel nature with units i nested in clusters j, for instance repeated measurements (units) nested in subjects (clusters) or subjects (units) nested in families (clusters). There will often be unobserved heterogeneity at the cluster level inducing dependence among the units, even after conditioning on covariates. In generalized linear mixed models (e.g. Breslow and Clayton, 1993) unobserved heterogeneity is modeled by including random effects $\eta_{mj}^{(2)}$ in the linear predictor,

$$g(\mu_{ij}) \; = \; \nu_{ij} = \; \underbrace{\mathbf{x}_{ij}'\boldsymbol{\beta}}_{\text{Fixed part}} \; + \; \underbrace{\sum_{m=1}^{M} \eta_{mj}^{(2)} z_{mij}^{(2)}}_{\text{Random part}}. \qquad (3)$$

Here, $\mu_{ij} \equiv \mathrm{E}[y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}^{(2)}, \boldsymbol{\eta}_j^{(2)}]$ where $\boldsymbol{\eta}_j^{(2)} = (\eta_{1j}^{(2)}, \cdots, \eta_{M,j}^{(2)})'$ are random effects varying at level 2 and $\mathbf{z}_{ij}^{(2)}$ corresponding covariates. Specifically, $\eta_{mj}^{(2)}$ is a random effect of covariate $z_{mij}^{(2)}$ for cluster j, a random intercept if $z_{mij}^{(2)} = 1$. It is typically assumed that the random effects are multivariate normal.

4.2 Extending GLMMs to GLLAMMs

Multilevel factor structures The basic idea of factor or IRT models is that one or more unobserved variables, latent traits or factors 'explain' the dependence between different observed measurements for a subject, in the sense that the measurements are conditionally independent given the factor(s).

A simple example of a unidimensional factor model is the two-parameter logistic item response model often used in ability testing. Examinees j answer test items i, $i = 1, \ldots, I$, giving responses y_{ij} equal to 1 if the answer is correct and 0 otherwise. The probability of a correct response is modelled as a function of the examinee's latent ability η_j,

$$\Pr(y_{ij} = 1|\eta_j) \; = \; \frac{\exp(\nu_{ij})}{1 + \exp(\nu_{ij})}, \qquad \nu_{ij} = \beta_i + \lambda_i \eta_j. \qquad (4)$$

The latent ability η_j is assumed to have a normal distribution, λ_i are factor loadings or discrimination parameters (with $\lambda_1 = 1$) signifying how well the items discriminate between examinees with different abilities, and $-\beta_i/\lambda_i$ are item 'difficulties'.

We can specify models of this form by extending the two-level generalized linear mixed model in (3) to allow each random effect to be multiplied not just by a single variable but by a linear combination of variables. To obtain the two-parameter logistic item response model, we stack the dichotomous responses y_{ij} into a single response vector and define dummy variables

$$d_{pi} = \left\{ \begin{array}{ll} 1 & \text{if } p=i \\ 0 & \text{otherwise} \end{array} \right.$$

The linear predictor of the item response model can then be written as

$$\nu_{ij} = \sum_p d_{pi}\beta_p + \eta_j \sum_p d_{pi}\lambda_p = \beta_i + \eta_j\lambda_i.$$

The linear predictor for a three-level multidimensional factor model can be expressed as

$$\nu_{ijk} = \underbrace{\mathbf{x}'_{ijk}\boldsymbol{\beta}}_{\text{Fixed part}} + \underbrace{\sum_{m_2=1}^{M_2} \eta^{(2)}_{m_2jk}\boldsymbol{\lambda}^{(2)\prime}_{m_2}\mathbf{z}^{(2)}_{m_2ijk}}_{\text{Level-2 random part}} + \underbrace{\sum_{m_3=1}^{M_3} \eta^{(3)}_{m_3k}\boldsymbol{\lambda}^{(3)\prime}_{m_3}\mathbf{z}^{(3)}_{m_3ijk}}_{\text{Level-3 random part}},$$

where $\mathbf{z}^{(2)}_{m_2ijk}$ and $\mathbf{z}^{(3)}_{m_3ijk}$ are vectors of dummy variables with corresponding vectors of factor loadings, $\boldsymbol{\lambda}^{(2)}_m$ and $\boldsymbol{\lambda}^{(3)}_m$. See Rabe-Hesketh, Skrondal and Pickles (2004a) for an application of a multilevel factor model with dichotomous responses.

Discrete latent variables The response model can be further generalized by allowing the latent variables η_j to have discrete distributions. This is useful if the level 2 units are believed to fall into a number of groups or 'latent classes' within which the latent variables do not vary.

If the number of latent classes, or masses, is chosen to maximize the likelihood the nonparametric maximum likelihood estimator (NPMLE) can be achieved (e.g. Rabe-Hesketh, Pickles and Skrondal, 2003), relaxing the assumption of multivariate normal latent variables.

Multilevel structural equations Continuous latent variables (random coefficients and/or factors) can be regressed on covariates (see Section 6) and other latent variables at the same or higher levels, generalizing conventional structural models to a multilevel setting. If the latent variables are discrete, the masses, component weights or latent class probabilities can depend on covariates via multinomial logit models. See Skrondal and Rabe-Hesketh (2004, Ch.4).

5   Composite links and exploded likelihoods in GLLAMMs

An outline is given of some extensions of GLLAMMs arising from plugging in linear predictors with latent variables from GLLAMMs into composite links and exploded likelihoods.

Discrete time frailty models If we let the linear predictor in (2) be $\nu_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \eta_j$ and use a logit link we can obtain a proportional odds model with frailty (see Skrondal and Rabe-Hesketh, 2004, Ch.12).

Item response models for ordinal items Letting the linear predictor in (2) be $\nu_{ij} = \beta_i + \lambda_i\eta_j$ as in the two parameter IRT model (4) and the thresholds be item-specific, we obtain Samejima's graded response model for ordinal items (see Skrondal and Rabe-Hesketh, 2004, Ch.10).

Unfolding or ideal point models In standard item response models the probability of a positive response for an item is a monotonic function of the latent trait η_j. This assumption may be violated for attitude items where respondents are asked to rate their agreement as 'disagree' or 'agree', or more generally in terms of $s = 1, \ldots, S$ ordered categories.

For instance, as sentiments favouring capital punishment increase from negative infinity, the probability of agreeing with the statement 'capital punishment seems wrong but is sometimes necessary' initially increases from 0, reaches a maximum when the latent trait is in the 'ambiguous' zone (at the 'ideal point') and then declines as the latent trait goes to infinity.

It has been argued (e.g. Roberts and Laughlin, 1996) that a respondent may give a particular rating of an attitude item for two reasons. Considering 'disagree', he can 'disagree from below' because his latent trait is below the position of the item or 'disagree from above' because it exceeds the position. These two possibilities can be expressed in terms of 'subjective ratings' z_{ij}; such that $z_{ij} = s$ if the respondent 'disagrees from below' and $z_{ij} = 2S+1-s$ if he 'disagrees from above'.

Since the z_{ij} are not observed, the probabilities of the observed rating y_{ij}, given the latent trait η_j, can be written as the sum of the probabilities of the two disjunct 'subjective ratings' corresponding to the observed rating. We propose using a cumulative model (2) for the subjective ratings

$$\Pr(y_{ij} = s|\eta_j) = \Pr(z_{ij} = s|\eta_j) + \Pr(z_{ij} = 2S+1-s|\eta_j) = \quad (5)$$
$$\left[g^{-1}(\nu_{ij} - \kappa_{s-1}) - g^{-1}(\nu_{ij} - \kappa_s) \right] + \left[g^{-1}(\nu_{ij} - \kappa_{2S-s}) - g^{-1}(\nu_{ij} - \kappa_{2S-s+1}) \right],$$

where $\nu_{ij} = \beta_i + \lambda_i\eta_j$ as in (4). For identification, the thresholds must be constrained as for instance $\kappa_s = -\kappa_{2S-s}$, $s = 1, \ldots, S$, and $\kappa_S = 0$.

Importantly, embedding the models in the GLLAMM framework produce a wide range of novel unfolding models. The latent trait can for instance be regressed on same or higher level latent variables and/or regressed on covariates as demonstrated in Section 6.

Small area estimation Rindskopf (1992) emphasizes that composite link functions are useful for modelling count data where some observed counts represent sums of counts for different groups of units, due to different kinds of missing or partially observed categorical variables. These ideas have been used by Tranmer et al. (2004) in random effects modeling and empirical Bayes prediction of area specific odds-ratios, for instance for the association between ethnicity and unemployment. They make use of one-way marginal tables from the census 'tabular output', e.g. unemployment rate and ethnic composition, in addition to borrowing strength from other areas as usual in empirical Bayes prediction.

Models combining discrete and continuous latent variables Latent class models can be specified by modeling the 'complete' data (including latent class membership) using log linear models. Since latent class membership is unknown, we must sum over the latent classes to obtain expected counts for the observed response patterns. For a two-class model with three dichotomous observed responses y_i, $i = 1, \ldots, 3$, a log-linear model with conditionally independent responses given latent class membership can be written as

$$\log \mu_{y_1 y_2 y_3 c} = \nu_{y_1 y_2 y_3 c} = \beta_0 + c\alpha_0 + \sum_i y_i \beta_i + \sum_i y_i c\alpha_i,$$

where $c = 0, 1$ is the latent class indicator, $\mu_{y_1 y_2 y_3 c}$ is the expected count for response pattern y_1, y_2, y_3 and latent class c, and β_p and α_p, $p = 0, \ldots, 3$ are parameters. The expected values $\mu_{y_1 y_2 y_3}$ of the observed counts are modeled as the sum of the class-specific expected counts,

$$\mu_{y_1 y_2 y_3} = \exp(\nu_{y_1 y_2 y_3 0}) + \exp(\nu_{y_1 y_2 y_3 1}).$$

Qu, Tan and Kutner (1996) include continuous random effects η_j within a latent class model to relax conditional independence among the responses given latent class membership. To incorporate subject-specific random effects in the model, we expand the data to obtain counts (0 or 1) for each response and latent class pattern for each subject j. The model can then be written as

$$\begin{aligned} \log \mu_{y_1 y_2 y_3 cj} = \nu_{y_1 y_2 y_3 cj} = \quad & \beta_0 + c\alpha_0 + \sum_i y_i \beta_i + \sum_i y_i c\alpha_i \\ & + \eta_j (\sum_i y_i (1 - c)\lambda_{i0} + \sum_i y_i c\lambda_{i1}), \end{aligned}$$

where η_j can be interpreted as subject j's propensity to have a '1' (e.g. score positively on a diagnostic test, have a symptom, be diagnosed by a rater), with item-specific effects λ_{i0} for those who are healthy and λ_{i1} for those who have the disease. Since the total count for each person j is fixed at 1, we can estimate the multinomial logit version of this model

$$\Pr(y_1 y_2 y_3 c | j) = \frac{\exp(\nu_{y_1 y_2 y_3 cj})}{\sum_{y_1 y_2 y_3 c} \exp(\nu_{y_1 y_2 y_3 cj})}$$

Again, we do not know c, so the likelihood contribution for subject j becomes

$$\Pr(y_1 y_2 y_3 | j) = \frac{\exp(\nu_{y_1 y_2 y_3 0j}) + \exp(\nu_{y_1 y_2 y_3 1j})}{\sum_{y_1 y_2 y_3 c} \exp(\nu_{y_1 y_2 y_3 cj})}.$$

This is a composite link model if each multinomial logit term is viewed as an inverse link. Note that this set-up makes it easy to relax conditional independence among pairs of items by including interaction effects of the form $\beta_{12} y_1 y_2$ in the linear predictors.

Item response models accommodating guessing If it is possible to guess the right answer of an 'item' in ability testing, as when multiple choice questions are used, the two-parameter logistic item response model in (4) is sometimes replaced by the three-parameter model

$$\Pr(y_{ij}=1|\eta_j) = c_i + (1-c_i)\frac{\exp(\nu_{ij})}{1+\exp(\nu_{ij})}.$$

The c_i are often called 'guessing parameters' and can be interpreted as the probability of a correct answer on item i for an examinee with ability minus infinity.

If we fix the guessing parameters to some common constant w, the response model can be expressed as a generalized linear model with a composite link

$$\Pr(y_{ij}=1|\eta_j) = wg_1^{-1}(1) + (1-w)g_2^{-1}(\nu_{ij}),$$

where g_1 is the identity link and g_2 is the logit link. If we let $\alpha_1 = w$ be a free parameter, we have a simple example of a bilinear composite link model.

The above kind of model (without latent variables) is said to have 'natural responsiveness' or 'nonzero background' in quantal response bioassay.

Log-normal random effects If the random effects distribution is skewed, we may want to specify a linear mixed model with log-normal random effects

$$\mu_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \exp(\eta_{1j}) + \exp(\eta_{2j})z_{ij},$$

which can be accomplished using the composite link

$$\mu_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \exp(\eta_{1j}) + \exp(\eta_{2j} + \log(z_{ij})).$$

This is also a useful way of conducting a sensitivity analysis of the conventional normality assumption for the random effects. Using the GLLAMM formulation, we can also have log-normal common factors.

If we use a bilinear composite link, we can include log-normal random effects in generalized linear mixed (and item response) models as well,

$$\mu_{ij} = h[\mathbf{x}'_{ij}\boldsymbol{\beta} + \exp(\eta_{1j}) + \exp(\eta_{2j} + \log(z_{ij}))].$$

Zero-inflated Poisson (ZIP) models The likelihood of ZIP models can be expressed using a combination of composite links and exploded likelihoods.

The ZIP model is a finite mixture model for counts where the population is assumed to consist of two components, a component $c=0$ where the count can only be zero and a component $c=1$ where the count has a Poisson distribution. The probability of belonging to the zero-count component is modelled as

$$\pi_{i0} = \frac{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}{1 + \exp(\mathbf{z}_i'\boldsymbol{\gamma})} \tag{6}$$

and the Poisson distribution for the other component is

$$\Pr(y_i = k | \mathbf{x}_i, c_i = 1) = \exp(-\mu_i)\mu_i^k/k!, \quad \mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}). \tag{7}$$

The probability of a non-zero count becomes

$$
\begin{aligned}
\Pr(y_i = k > 0 | \mathbf{z}_i, \mathbf{x}_i) &= \Pr(y_i = k > 0, c_i = 1) = (1 - \pi_{i0})\exp(-\mu_i)\mu_i^k/k! \\
&= \left(\frac{1}{1 + \exp(\mathbf{z}_i'\boldsymbol{\gamma})} \right) \left[\exp(-\mu_i)\mu_i^k/k! \right]
\end{aligned}
$$

and the probability of a zero count

$$
\begin{aligned}
\Pr(y_i = 0 | \mathbf{z}_i, \mathbf{x}_i) &= \Pr(y_i = 0, c_i = 0 | \mathbf{z}_i, \mathbf{x}_i) + \Pr(y_i = 0, c_i = 1 | \mathbf{z}_i, \mathbf{x}_i) \\
&= \pi_{i0} + (1 - \pi_{i0})\exp(-\mu_i) \\
&= \left(\frac{1}{1 + \exp(\mathbf{z}_i'\boldsymbol{\gamma})} \right) \left[\exp(\mathbf{z}_i'\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta})) \right].
\end{aligned}
$$

For a non-zero count, the probability is the product of the probability of 0 in a logistic regression model with linear predictor $\mathbf{z}_i'\boldsymbol{\gamma}$ and the Poisson probability of a count k with a log link and linear predictor $\mathbf{x}_i'\boldsymbol{\beta}$. Therefore, for non-zero counts, we obtain the correct likelihood by creating two responses, 0 and k and specifying a mixed response (logistic and Poisson) model.

For a zero count, we again create a 0 response, modelled as a logistic regression, for the first term. For the second term, we specify a composite link,

$$\left[\exp(\mathbf{z}_i'\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta})) \right] = g_1^{-1}(\mathbf{z}_i'\boldsymbol{\gamma}) + g_2^{-1}(\mathbf{x}_i'\boldsymbol{\beta}),$$

where g_1 is the log link and g_2 the log-log link. If we create a 1 response and specify a Bernoulli distribution with this composite link, we obtain the required term.

This set-up also makes it fairly straightforward to include random effects in ZIP models to capture dependence induced by clustered data. For instance, in modeling the number of alcoholic drinks consumed by respondents nested in regions, we could include region-specific random effects in both (6) and (7) to model variations in the prevalence of non-drinking and in the amount consumed among drinkers, with possible correlations between these random effects.

6   Unfolding attitudes to female work participation

In the 1988 and 2002 General Social Surveys respondents in the USA were presented with the following attitude statements regarding female work participation:

[famhapp] A woman and her family will all be happier if she goes to work

[twoincs] Both the husband and wife should contribute to the family income

[warmrel]: A working mother can establish just as warm and secure a relationship with her children as a woman who does not work

[jobindep] Having a job is the best way for a woman to be an independent person

[housewrk] Being a housewife is just as fulfilling as working for pay

[homekid] A job is alright, but what most women really want is a home and children

[famsuff] All in all, family life suffers when the woman has a full-time job

[kidsuff] A pre-school child is likely to suffer if his or her mother works

[hubbywrk] A husband's job is to earn money; a wife's job is to look after the home

The respondents rated each statement as either 'disagree completely' (1), 'disagree' (2), 'agree somewhat' (3), 'agree' (4), or 'agree completely' (5). In 2002, the 'disagree completely' and 'disagree' response options were collapsed into a single 'disagree' option.

We use the unfolding model proposed in Section 5, with g as scaled probit links with item-specific scale parameters σ_i (estimated on the log-scale),

$$g^{-1}(\nu_{ijs}) \;=\; \Phi^{-1}\left(\frac{\beta_i + \lambda_i \eta_j - \kappa_s}{\sigma_i}\right).$$

In 2002, the composite link for 'disagree' is the sum of the composite links for 'disagree' and 'disagree completely'.

To investigate if sentiments in favour of female work participation η_j (loosely referred to as 'feminism') have changed from 1988 to 2002, we specify the structural model

$$\eta_j \;=\; \gamma_1 w_j + \zeta_j, \qquad \zeta_j \sim \mathrm{N}(0, \psi),$$

where w_j is a dummy variable for year being [2002].

Maximum likelihood estimates based on data from 1462 respondents are given in Table 1 where the items have been ordered from the most positive to the most negative according to their estimated scale values $\widehat{\beta}_i$. Since the magnitude of $\widehat{\gamma}_1$ is negligible, mean 'feminism' does not appear to have changed.

TABLE 1. Estimates for scaled probit unfolding model

| | Item parameters | | | | | |
Item i	β_i Est	SE	λ_i Est	SE	$\ln \sigma_i$ Est	SE
[famhapp]	-2.32	0.08	0.30	0.04	-0.24	0.05
[twoincs]	-1.60	0.07	0.29	0.05	-0.06	0.05
[warmrel]	-0.99	0.07	1	–	0	–
[jobindep]	-0.27	0.14	1.15	0.15	0.64	0.05
[housewrk]	1.29	0.08	0.54	0.08	0.22	0.06
[homekid]	2.11	0.07	0.76	0.06	-0.06	0.04
[famsuff]	2.19	0.08	1.43	0.09	-0.29	0.05
[kidsuff]	2.24	0.08	1.49	0.09	-0.46	0.06
[hubbywrk]	2.42	0.09	1.14	0.09	-0.11	0.05

Thresholds $-\kappa_s = \kappa_{2S-s}$		
s (categories)	Est	SE
1 ('disagree completely'/'disagree')	3.43	0.11
2 ('disagree'/'agree somewhat')	2.36	0.08
3 ('agree somewhat'/'agree')	1.67	0.06
4 ('agree/'agree completely')	0.72	0.03

Latent trait regression		
	Est	SE
[2002] γ_1	-0.04	0.04
Variance ψ	0.62	0.08

Following Roberts and Laughlin (1996) we assess model fit graphically. First, we estimate the position or 'dominance' $\widetilde{\nu}_{ij}$ of respondent j relative to item i (how much more 'feminist' the respondent is than the item) by plugging in the empirical Bayes prediction $\widetilde{\eta}_j$ of the latent trait and the parameter estimates into the linear predictor. Substituting this into the unfolding model, we obtain the expected response category for each person-item pair. Grouping the $\widetilde{\nu}_{ij}$ into approximately homogeneous groups of size 30 for each item and plotting the corresponding average observed and expected frequencies versus the average $\widetilde{\nu}_{ij}$ for each item gives Figure 1. Our unfolding model appears to fit quite well.

Although the expected response takes the form of a single-peaked function consistent with an unfolding process when all items are considered together, none of the individual items exhibit single-peaked behaviour with the possible exception of [jobindep]. Using conventional item response models that assume monotonicity might therefore be appropriate if either (1) reversing the coding of the appropriate items can be based on a priori information or (2) the model accommodates negative factor loadings.
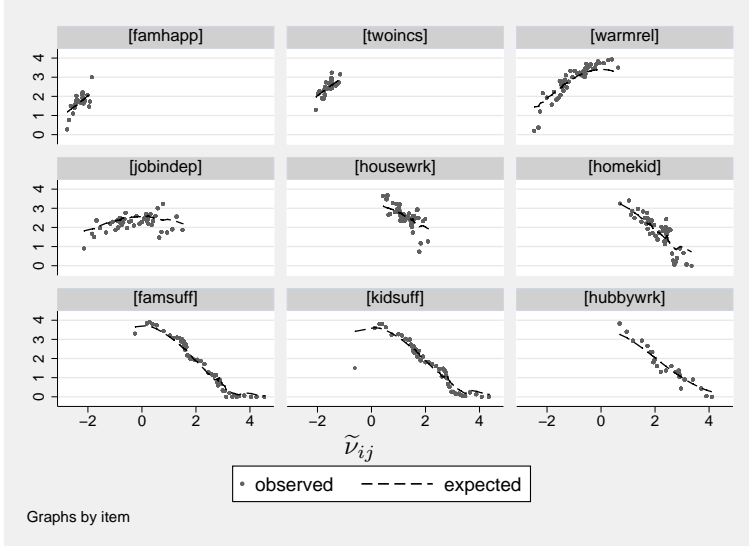
FIGURE 1. Mean expected and observed responses as a function of 'dominance' $\widetilde{\nu}_{ij}$ of person j over item i

7   Conclusions

Although simple to implement, composite links and exploded likelihoods have been demonstrated to be remarkably powerful tools for specifying novel GLLAMMs. Indeed, we do not purport to exhaust potential applications in this paper.

A further useful extension would be to generalize the traditional composite links suggested by Thompson and Baker (1981) to accommodate products of inverse links. A simple variant is of the form

$$\mu_i \;=\; \sum_r \alpha_r \prod_t g_{rt}^{-1}(\nu_{irt}).$$

A composite link with products can be used for additive relative risk models with random effects. The risk or rate parameter μ_{ij} in the Poisson distribution is specified as

$$\mu_{ij} \;=\; \exp(\beta_0 + \eta_j)[1 + \mathbf{x}'_{ij}\boldsymbol{\beta}],$$

where $\mathbf{x}_{ij}$ does not include a 1 and $\boldsymbol{\beta}$ correspondingly not a constant. Note that the baseline risk when $\mathbf{x}_{ij} = \mathbf{0}$ becomes $\exp(\beta_0 + \eta_j) > 0$. It follows that the 'relative risk' RR_{ij}, the risk when the covariate vector is $\mathbf{x}_{ij}$ relative to the baseline risk, is

$$\mathrm{RR}_{ij} \;=\; 1 + \mathbf{x}'_{ij}\boldsymbol{\beta},$$

an additive function of the covariates.

Maximum likelihood estimation and of GLLAMMs and empirical Bayes prediction using adaptive quadrature (e.g. Rabe-Hesketh, Skrondal and Pickles, 2004b) are implemented in the `gllamm` software running in `Stata`. See `http://www.gllamm.org` for further information.

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.

Qu, Y., Tan, M., and Kutner, M.H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, **52**, 797-810

Rabe-Hesketh, S., Pickles, A., and Skrondal, A. (2003). Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statistical Modelling*, **3**, 215-232.

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2004a). Generalized multilevel structural equation modeling. *Psychometrika*, in press.

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2004b). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, in press.

Rindskopf, D. (1992). A general approach to categorical data analysis with missing data, using generalized linear models with composite links. *Psychometrika*, **57**, 29-42.

Roberts, J.S., and Laughlin, J.E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement*, **20**, 231-255.

Skrondal, A., Rabe-Hesketh, S. (2003). Multilevel logistic regression for polytomous data and rankings. *Psychometrika*, **68**, 267-287.

Skrondal, A., Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling. Multilevel, Longitudinal and Structural Equation Models*. Boca Raton, FL: Chapman & Hall/CRC.

Thompson, R., and Baker, R.J. (1981). Composite link functions in generalized linear models. *Journal of the Royal Statistical Society, Series C*, **30**, 125-131.

Tranmer, M., Pickles, A., Fieldhouse, E., et al. (2004). The case for small area microdata. *Journal of the Royal Statistical Society, Series A*, in press.

Statistical Analysis of Replicated Microarray Time Series Data

Terry Speed[1] and Yu Chuan Tai[1]

[1] Department of Statistics and Program in Biostatistics University of California at Berkeley

Abstract: We describe a one-sample multivariate empirical Bayes statistic (the MB statistic) to select differentially expressed genes from replicated microarray time course experiments. We do this by testing the null hypothesis that the expectation of a k-vector of a gene's expression levels is a multiple of 1_k, the vector of k 1s. The importance of moderation in this context is explained. Together with the MB statistic we have the one-sample $\tilde{T}^2$ statistic, a variant of the one-sample Hotelling T^2. Both the MB statistic and $\tilde{T}^2$ can be used to rank genes in the order of evidence of nonconstancy, incorporating the correlation structure among time point samples and the replication. In a simulation study we show that the MB statistic and $\tilde{T}^2$ statistic achieve the smallest number of false positives and false negatives, and perform slightly better than the one-sample moderated Hotelling T^2 statistic. Several special and limiting cases of the MB statistic are derived, and two-sample versions described. Finally, we illustrate the use of these statistics in two microarray time course studies.

Oral Sessions

Model Selection for Regression Analyses with Missing Data

M. Aerts[1], N. Hens[1] and G. Molenberghs[1]

[1] Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium

Abstract: The Akaiki Information Criterion, AIC, is one of the leading selection methods for regression models. In case of partially missing covariates with missingness probability depending on the response, regression estimates based on the so-called complete cases are known to be biased. In this contribution it is shown that model selection using AIC-values based on the complete cases can lead to the choice of wrong or less optimal models. In analogy with the weighted Horvitz-Thompson estimator, we propose a weighted version of AIC. It is shown that this weighted AIC criterion improves model choices.

1 Introduction

Let $(x_1, z_1, y_1), ..., (x_n, z_n, y_n)$ be a sample where y denotes a response variable and x and z covariate variables. Here we focus on the case that, for a fixed value of x and z, the response y is normally distributed with variance σ^2. Suppose we want to select an optimal model from a set of K candidate models for the mean function $\mu(x, z) = E(y|x, z)$. A well-established method is selecting the model k which minimizes the AIC criterion (Akaike 1973, Linhart and Zucchini 1986, Burnham and Anderson 1998, Hurvich and Tsai 1989):

$$AIC = -2 \log(\text{likelihood of model } k) + 2 \times (\text{\# parameters of model } k), \tag{1}$$

where the likelihood is evaluated at the corresponding ML-estimator. For a normal error structure, this simplifies to (ignoring some constant terms, not depending on k):

$$AIC = n \log \hat{\sigma}_k^2 + 2p_k, \tag{2}$$

where $\hat{\sigma}_k^2$ is the ML variance estimator based on model k and p_k is the number of regression coefficients in model k.

In a missing data context, covariate x or response y may be missing. We assume z is always observed. Let $\delta_i = 1$ if the ith observation is completely observed and $\delta_i = 0$ otherwise. Furthermore, let the selection probabilities

$\pi_i = P(\delta_i = 1|y_i, x_i, z_i)$ reflect the missing at random (MAR) missingness mechanism (Rubin 1976). So, $\pi_i = P(\delta_i = 1|y_i, z_i)$ in the missing covariate case and $\pi_i = P(\delta_i = 1|x_i, z_i)$ in case the response y is subject to missingness. For missing covariate data, Flanders and Greenland (1991) and Zhao and Lipsitz (1992) suggested a weighted estimator in the spirit of Horvitz and Thompson (1952), based on the weighted likelihood or weighted least squares criterion for the complete cases (CC) with weights equal to $1/\hat{\pi}_i$, where $\hat{\pi}_i$ is an appropriate estimator for the selection probabilities π_i. Wang et al. (1997) proposed to use a nonparametric kernel smoother to estimate the selection probabilities while fitting the regression curve with a parametric model and Wang et al. (1998) proposed a weighted local linear estimator for $\mu(x)$ while using local linear estimates for $\pi(y_i)$.

Model selection for incomplete data has not received much attention in the literature. Cavanaugh and Shumway (1998) derived and investigated a variant of AIC motivated by the same principle as the 'predictive divergence of incomplete observations'. Hens, Aerts and Molenberghs (2004) proposed modifications of several model selection criteria using weighting likelihood ideas and compared it to "model selection after imputation" methods. A similar weighted Akaiki information criterion in the context of robust model selection and robust regression models has been proposed by Agostinelli (2002).

2   Modified AIC criterion

We focus on the weighted AIC criterion applied to normal response data as described in the previous section. Weighting in (2) each complete case contribution to the loglikelihood with weight $1/\hat{\pi}_i$ leads to the criterion

$$AIC_W = (\sum_{i=1}^{n} \delta_i/\hat{\pi}_i) \log \hat{\sigma}_{W,k}^2 + 2p_k \qquad (3)$$

where $\hat{\sigma}_{W,k}^2$ is the ML variance estimator based on the weighted (normal) likelihood.

3   Unknown weights

In some settings (e.g. a two-stage design), the selection probabilities are known and do not have to be estimated. In many missing data problems, however, the unknown weights π_i, which can be considered as nuisance parameters, have to be estimated. This estimator has to be consistent, otherwise it will adversely affect the model selection procedure. So if we estimate π_i with a parametric model, we are faced with an additional model selection problem. Hens, Aerts and Molenberghs (2004) suggest the use of a nonparametric estimator, e.g. a kernel smoother as used in Wang et

al. (1998). In the next section we illustrate the applicability of the method in a small simulation study.

4 Simulation Study and Discussion

Observations for a continuous explanatory variable X are generated from a uniform distribution on the interval $[0, 10]$, Z observations are generated from a Bernoulli distribution with probability 0.50. Conditionally upon X, Y observations are generated from a normal distribution with mean $\mu(x) = -3 + 3x + 5x^2$ and variance $\sigma^2 = \exp(5)$. X observations are then turned into 'missing' with conditional probability $\pi(x) = [1 + \exp\{1 - 0.009(y - 300)\}]^{-1}$. We generated 1000 different samples $\{Y_i, i = 1, \ldots, n\}$ with a fixed design $\{x_i, z_i, i = 1 \ldots, n\}$ of sample size $n = 100$. For each sample, 8 different regression models were fit, i.e. all submodels of $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 Z + \beta_4 XZ$.

Model Method	1	X	Z	X, X^2	X, Z	$X, X^2,$ Z	$X, Z,$ XZ	$X, X^2,$ Z, XZ
ALL	0	125	0	647	30	128	13	57
CC	0	340	0	432	71	75	38	44
TW	0	197	0	366	74	116	69	178
EW	0	269	0	422	73	97	52	87
E2	0	220	0	396	78	103	66	137

TABLE 1. Simulation study with 8 candidate models: number of AIC selected models

Method	Correct	Incorrect
ALL	832	168
CC	551	449
TW	660	340
EW	606	394
EW2	636	364

TABLE 2. Simulation study with correctly and incorrectly classified models: number of AIC selected models

Table 1 shows, for each candidate model, the number of times it is has been selected as best model by the AIC criterion (2) or (3), for 5 different methods: ALL stands for an unweighted analysis based on all data (as if no data were missing); CC for an unweighted analysis on the complete cases only (excluding the observations with a missing X-value); TW for a

weighted analysis with true known missingness probabilities $\pi(x)$; EW for a weighted analysis with kernel estimated probabilities $\pi(x)$ using a fixed bandwidth and finally, EW2 for a weighted analysis with kernel estimated probabilities using a cross-validation data-driven choice of the smoothing parameter.

A comparison of the first two rows shows the effect of ignoring the missingness by using an unweighted AIC criterion on the complete cases. The weighted criterion (3) improves the selection of correct models, as shown in the last three rows of Table 1 and Table 2. In Table 2, all more complex models containing the true model as a submodel are collapsed in a category "correct model".

The last two lines illustrates the importance of using a data-driven smoothing parameter, when estimating the missingness probabilities $\pi(x)$.

References

Agostinelli, C. (2002). Robust model selection in regression via weighted likelihood methodology. *Stat. & Prob. Letters*, **56**, 289–300.

Akaiki, H. (1973). Information theory as an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*. Petrov, B.N., and Csaki, F. (eds.), Akademiai Kiado, 267–281.

Burnham, K.P. and Anderson, D.R. (1998). *Model Selection and Inference*. New York: Springer-Verlag.

Cavanaugh, J.E. and Shumway, R.H. (1998). An Akaiki information criterion for model selection in the presence of incomplete data. *J. Statist. Plan. Inf.*, **67**, 45–65.

Flanders, W.D. and Greenland, S. (1991). Analytic methods for two-stage case-control studies and other stratified designs. *Stat. in Med.*, **10**, 739–747.

Hens, N., Aerts, M. and Molenberghs, G. (2004). Regression model selection for incomplete and non-random samples. Technical Report.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, **47**, 663–685.

Hurvich, C.M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.

Linhart, H. and Zucchini, W. (1986). *Model Selection*. New York: Wiley.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

Wang, C.Y., Wang, S., Gutierrez, R.G., and Carroll, R.J. (1998). Local linear regression for generalized linear models with missing data. *Ann. Statist.*, **26**, 1028–1050.

Wang, C.Y., Wang, S., Zhao, L-P., Ou, S-T. (1997). Weighted semiparametric estimation in regression analysis with missing covariate data. *J. Amer. Statist. Assoc.*, **92**, 512 –525.

Zhao, L.P. and Lipsitz, S. (1992). Design and analysis of two-stage studies. *Stat. in Medicine*, **11**, 769–782.

A Computationally Tractable Multivariate Random Effects Model for Clustered Binary Data

Brent A. Coull[1], E. Andres Houseman[1], Rebecca A. Betensky[1]

[1] Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA, email: **bcoull@hsph.harvard.edu**

Abstract: We consider a multivariate random effects model for clustered binary data that is useful when interest focuses on the association structure among clustered observations. Based on a vector of gamma random effects and a complementary log-log link function, the model yields a likelihood that has closed-form, making a frequentist approach to model fitting straightforward. We consider the interpretation and identifiability of the model parameters, and use the proposed model to analyze binary time series data from an arthritis clinical trial.

Keywords: complementary log-log link; binary time series; generalized linear mixed model; multivariate gamma.

1 Introduction

Use of generalized linear mixed models (GLMM; Breslow and Clayton 1993) has become a popular approach to modeling correlated discrete data. The models account for correlation among clustered observations by including random effects in the linear predictor component of the model. Although GLMM model fitting is typically complex, standard random intercept and random intercept and slope models can now be routinely implemented in such commercial software packages as SAS, Stata, and Splus/R.

While in many applications the nature of dependence between clustered responses is a nuisance, in some scientific settings interest focuses primarily on the association structure among clustered observations. Examples include studies focusing on serially correlated observations (e.g. Fitzmaurice and Lipsitz 1995) and familial aggregation of disease (Betensky and Whittemore 1996). A disadvantage of standard GLMMs in these instances is their inability to handle complex dependence structures among clustered responses. Several authors have proposed adding additional random effects to flexibly model more complicated association structures (e.g. Diggle et al. 2002, Section 11.4.2). These additional random effects, however, add a layer of complexity to model fitting.

We consider a multivariate random effects model for clustered binary data that is useful when interest focuses on the association structure among

clustered observations. The model represents a multivariate random effects extension of a model proposed by Conaway (1990). Based on a vector of gamma random effects and a complementary log-log link function, the proposed model yields a marginal likelihood that has closed-form, making computationally intensive numerical integration or Monte Carlo sampling unnecessary. As a result, model fitting via maximum likelihood is computationally simple. In addition, as we discuss further in Section 3, a closed-form likelihood allows the user to check model identifiability relatively easily.

2   Model

Let the vector $\mathbf{Z} = (Z_1, \ldots, Z_p)^\mathsf{T}$ be multivariate gamma as defined by Henderson and Shimakura (2003); that is, for suitable choice of matrix $\mathbf{C} = ((c_{ij}))$, $\mathbf{Z}$ has Laplace transform

$$\mathcal{L} = E\left\{\exp\left(-\mathbf{u}^\mathsf{T}\mathbf{Z}\right)\right\} = |\mathbf{I} + \zeta\mathbf{C}\mathrm{diag}(\mathbf{u})|^{-1/\zeta}, \tag{1}$$

for all $\zeta > 0$. Marginally, $Z_j \sim \mathrm{Gamma}(1/\zeta, 1/\zeta)$, $j = 1, \ldots, p$, with correlation matrix describing the association among gamma variables equal to $\mathbf{R}$ with elements $r_{jk} = c_{jk}^2$. We denote this multivariate distribution $\mathbf{Z} \sim MG(\zeta, \mathbf{C})$.

Now, let Y_{ij} denote binary response j, $j = 1, \ldots, n_i$, in cluster i, $i = 1, \ldots, N$. Let $\theta_{ij} = \log(Z_{ij})$ be a random effect corresponding to Y_{ij}, and consider the GLMM

$$\ln\left\{-\ln\left[\mathrm{E}\left(Y_{ij}|\mathbf{Z}_i\right)\right]\right\} = \theta_{ij} + \mathbf{x}_{ij}^\mathsf{T}\boldsymbol{\beta}, \tag{2}$$

where $\mathbf{Z}_i \overset{iid}{\sim} MG(\zeta, \mathbf{C})$ and $\boldsymbol{\beta}$ is a $k \times 1$ vector of fixed effects. In this framework, ζ is an overdispersion parameter, the interpretation of which we address in detail in Section 3. Interest typically focuses on both the fixed effects $\boldsymbol{\beta}$ and the matrix $\mathbf{C}$, often parameterized as a known function of a smaller number of variance components $\boldsymbol{\rho}$.

In order to derive the joint probability $P(Y_{i1} = y_1, Y_{i2} = y_2, \ldots, Y_{in_i} = y_{n_i})$, we use the method of Conaway (1990) of first computing marginal probabilities in the 2^{n_i} table that cross-classifies the binary responses in a given cluster, and subsequently transforming these marginal probabilities back to the joint probabilities of interest. Suppressing the i notation, let T be a subset of the indices $\{1, 2, \ldots, n\}$, and define

$$\pi_T^* = \int \prod_{j \in T} P(Y_j = 1|\mathbf{Z})f(\mathbf{Z})d\mathbf{Z}. \tag{3}$$

Under model (2), these probabilities have closed form:

$$\pi_T^* = \int \exp\left\{-\sum_{j \in T} Z_{ij}\exp\left(\mathbf{x}_{ij}^\mathsf{T}\boldsymbol{\beta}\right)\right\} f(\mathbf{Z})d\mathbf{Z}$$

$$=\quad |\mathbf{I} + \zeta\mathbf{C}\mathrm{diag}(\mathbf{u})|^{-1/\zeta},$$

where the jth element of $\mathbf{u}$ equals $\exp(\mathbf{x}_j^\mathsf{T}\boldsymbol{\beta})$ if $j \in T$ and 0 otherwise. Thus, only changes in the elements of $\mathbf{u}$ are necessary to reflect differences among specific π_T^*. If $\boldsymbol{\pi}^*$ is the collection of all such marginal probabilities π_T^*, then the vector of probabilities defining the joint distribution of $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is a known linear transformation of $\boldsymbol{\pi}^*$, yielding a marginal likelihood having closed-form. We maximize the corresponding log likelihood with respect to $(\boldsymbol{\beta}, \boldsymbol{\rho}, \zeta)$ using the optimization function `optim` in the R software package, and base inference on the inverse Hessian matrix evaluated at the maximum likelihood estimates.

3   Parameter Identifiability

We now discuss the identifiability and interpretation of the parameters ζ and ρ in the complementary log-log – multivariate gamma model. For concreteness, we focus on the first order-autoregressive correlation structure $c_{ik} = \rho^{|t_i - t_k|}$, although similar reasoning applies for other correlation structures such as the compound symmetric structure $c_{ik} = \rho$.

To understand the model parameters, it is instructive to consider special cases of the model with parameters held fixed at specific values. When $\rho = 0$, the individual gamma random effects, and hence the binary responses, are independent. In this case, the data are unclustered, and the overdispersion parameter ζ is unidentifiable in the presence of a mean model for π_{ij}. In contrast, the special case of the model with $\rho = 1.0$ corresponds to the simple random intercept model proposed by Conaway (1990). In this case, ζ represents the variance component for the random intercepts in the model, and is clearly identifiable. Thus, identifiability of the model parameters depends on the strength of the association among clustered responses, with the model being weakly identifiable for a wide range of ρ values within the two extremes. Simulations confirm these likelihood properties, and suggest that all model parameters are estimable when ρ is greater than approximately 0.90. The above identifiability considerations are not unique to the complementary log-log model considered here, but apply to other multivariate random effects models as well.

To address cases of weak identifiability in a frequentist approach to fitting model (2), we propose first fitting the model fixing the overdispersion parameter ζ to be 1.0. Simulations suggest that this approach results in well-identified parameters in this multivariate gamma setting. If the estimated correlation ρ under this constraint is not large, the overdispersion parameter ζ is likely not identifiable from the data. In cases in which there is strong association among outcomes, we propose then fitting the unconstrained model and estimating ζ. The closed-form likelihood enables the

user to check model identifiability relatively easily by inspecting likelihood contours and the information matrix of the resulting parameter estimates. See Coull, Houseman, and Betensky (2004) for further details.

4   Example: Binary Time Series Data

We apply the proposed model to binary time series data from an arthritis clinical trial. For each of $N = 51$ subjects, the data consist of at most five unequally spaced binary self-assessment measurements of arthritis, with this outcome equaling 0 if "poor" and 1 if "good". Patients were randomized to one of two drug treatments, placebo or auranofin. Patients had self-assessments taken at week 0 and week 1 prior to randomization, and at weeks 5, 9, and 13 post-randomization. Interest focuses on the effect of drug treatment, while controlling for gender, age at week 0, and time (in weeks). Of the 51 subjects, 14 (27%) have some missing responses.

We analyze the data with the main effects model

$$\ln\left\{-\ln\left[\mathrm{E}\left(Y_{ij}|\mathbf{Z}_i\right)\right]\right\} = \theta_{ij} + \beta_0 + \beta_1 \mathrm{Age}_i + \beta_2 \mathrm{Time}_{ij} + \beta_3 \mathrm{Drug}_{ij} + \mathrm{Gender}_i, \tag{4}$$

assuming the exponential correlation structure $c_{jk} = \rho^{|t_j - t_k|}$ for the multivariate gamma random effects. In view of the identifiability considerations outlined in Section 3, we run a preliminary analysis constraining the overdispersion parameter to be $\zeta = 1.0$. The estimated serial correlation parameter in this case is $\widehat{\rho} = 1.0$, indicating that the association is strong in this setting.

The unconstrained fit yields $\widehat{\zeta} = 3.29$, which is far from 1.0. In addition, a comparison the maximum likelihoods suggests that the unconstrained model fits significantly better than the constrained model, although a condition number of 1.29×10^4 for the Hessian matrix suggests that the likelihood is somewhat flat in the (β_0, ζ) direction. The estimate $\widehat{\rho} = 0.978$ again suggests strong correlation among adjacent outcomes. Because the model contains a continuous covariate, goodness-of-fit measures for contingency tables do not directly apply to this model. However, a goodness-of-fit test applied to the one way table classifying subjects according to their number of "good" self-assessments suggests that the model fits well ($p = 0.82$). We re-fit the model after dropping non-significant terms Age, Time, and Gender. Under this simpler model, the estimated drug effect corresponds to a log odds ratio of 1.98, which, as expected, is larger than the GEE estimate of 1.45 obtained by Fitzmaurice and Lipsitz (1995).

5   Discussion

In this article we have proposed a new multivariate random effects model for clustered binary observations. The model provides flexibility in modeling the association structure among observations, and maximum likelihood

inference is computationally straightforward. In the clinical trial example considered in the previous section, the model provides a likelihood-based approach to analyzing serially correlated binary responses.

As noted in Section 3, such multivariate random effects models for binary responses can be over-parameterized for some data configurations. We have proposed a careful inspection of the likelihood surface, via both likelihood plots and calculation of the condition number of the Hessian matrix evaluated at the MLE's. We view the ability to conduct such inspections using the proposed model one of its advantages over existing formulations for which closed-form expressions for the marginal likelihood do not exist.

References

Betensky, R. A. and Whittemore, A. S. (1996). An analysis of correlated multivariate binary data: Application to familial cancers of the ovary and breast. *Journal of the Royal Statistical Society, Series C* **45**, 411–429.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9-25.

Conaway, M. R. (1990). A random effects model for binary data. *Biometrics* **46**, 317-328.

Coull, B.A., Houseman, E.A., and Betensky, R.A. (2004). A computationally tractable multivariate random effects model for clustered binary data. Unpublished Technical Report.

Diggle, P.J., Heagerty, P., Liang, K-Y., Zeger, S.L. (2002). *Analysis of Longitudinal Data*. 2nd Ed. Oxford: Clarendon Press.

Fitzmaurice, G.M., Lipsitz, S.R. (1995). A model for binary time-series data with serial odds ratio patterns. *Applied Statistics* **44**, 51–61.

Henderson, R., Shimakura, S. (2003). A serially correlated gamma frailty model for longitudinal count data. *Biometrika* **90**, 355–366.

Localizing Clusters in Space-Time Point Process Data

Renato Assunção[1], Andréa Tavares [1] and Martin Kulldorff[2]

[1] Departamento de Estatística, Universidade Federal de Minas Gerais, 31270-901 Belo Horizonte, MG, Brazil. assuncao@est.ufmg.br

[2] Department of Ambulatory Care and Prevention, Harvard Medical School and Harvard Pilgrim Health Care, Boston USA

Abstract: Space-time interaction occurs in a point process when there are space-time clusters not explained by neither the purely spatial nor the purely temporal clustering. Knox, and others after him, have proposed tests to determine if there is space-time interaction as a general phenomena in a data set. These methods have been widely used in epidemiology, ecology and other fields. Sometimes it is also of interest to know the specific location of space-time interaction clusters. In this paper, we propose a new statistical method for the detection and inference of local space-time interaction clusters. It is based on scanning the three-dimensional space with a score test statistic under the null hypothesis that the point process is an inhomogeneous Poisson point process with space and time separable first order intensity. The method is illustrated using crime statistics from Belo Horizonte, Brazil, with the goal of finding space-time clusters of robberies and homicides not explained by purely spatial and purely temporal patterns.

Keywords: spatial statistics; point process; point pattern; scan statistic; score test; crime statistics.

Introduction

Crime varies substantially on space and time and separate analysis of these dimensions are often carried out. Less common is the simultaneous analysis of both dimensions aiming at, for example, finding evidence for the presence of any space-time clusters not explained by the baseline geographical and temporal variation. These are denoted as space-time interaction clusters. Knox (1964) proposed a test for space-time interaction that has been incorporated into various spatial statistical software and which is widely used in epidemiology, ecology and criminology. Mantel (1967), among other authors, proposed other space-time interaction tests. As with Knox test, these all have in common that they are general tests evaluating whether there is space-time interaction throughout the data, without pinpointing the location of specific clusters. That is very useful if we for example want to determine whether a particular disease may be infective of not, or if one

is interested in the general patterns of crime in order to understand sociological and behavioral aspects of criminal behavior. They are less useful for a police department wanting to know where and when to allocate their resources most effectively, or a public health official wanting to know the time and location of a disease outbreak, both of which requires knowledge of the space and time parameters of specific clusters.

Therefore, it is useful to differentiate two different types of alternatives to the null hypothesis of no space-time interaction. One of them focus on space-time clustering occurring throughout the map, either due to many small clusters of slightly larger than average incidence rate or many weakly interacting clusters of events. The other focus on situations where one or a few localized space-time clusters will have a substantially higher incidence rate, or where there is strong interaction between a subset of the events. For this second type of alternative, it is of interest to detect the location and time of specific clusters.

In this paper, we are interested in the first type of alternatives to lack of space-time clustering. We present our new space-time cluster detection test for space-time point processes in the next section. It uses a scan statistic approach and it does not requires risk population information or critical thresholds on space and time. Furthermore, our proposal is able to identify the specific space-time regions leading to rejection of the null hypothesis. We apply the methodology to three crime data sets. We conclude in Section 5 with a discussion on the potential value and limitations of our results for applications.

1   The new space-time test

In this section, we describe briefly the new test. Assume that we observe random point events generated by a Poisson point process in a space-time region $\mathcal{A} = A \times [0, \tau]$, where A is a bi-dimensional polygon. Given the observed events, the log-likelihood is equal to

$$l = \sum_{i=1}^{n} \log \lambda \left(x_i, y_i, t_i \right) - \int_{\mathcal{A}} \lambda(x, y, t) dx dy dt$$

The null hypothesis of no space-time interaction implies that the intensity function is equal to

$$H_0 : \lambda(x, y, t) = \lambda_S(x, y) \lambda_T(t)$$

Let $C = C_S \times C_T$ be a fixed and arbitrary space-time cylinder with C_S being a convex region in A and C_T a time interval. Consider a local alternative $H_{C,\epsilon}$ to H_0 given by

$$H_{C,\epsilon} : \lambda(x, y, t) = \lambda_S(x, y) \lambda_T(t) \left(1 + \epsilon I_C(x, y, t) \right)$$

where $\epsilon > 0$ and I_C is the indicator function that $(x, y, t) \in C$. For this hypothesis pair, the score test statistic is given by

$$\frac{\partial l}{\partial \epsilon}|_{\epsilon=0} = N(C) - \int_{C_S} \lambda_S(x, y) dx dy \times \int_{C_T} \lambda_T(t) dt \tag{1}$$

which can be estimated by

$$N(C) - \frac{N(C_S \times [0, T]) \, N(A \times C_T)}{N(A \times [0, T])} \tag{2}$$

Since $N(C)$ is a Poisson random variable, we propose to use

$$U_C = \frac{N(C) - N(C_S \times [0, T]) \, N(A \times C_T)/N(A \times [0, T])}{\sqrt{N(C_S \times [0, T]) \, N(A \times C_T)/N(A \times [0, T])}} \tag{3}$$

as a test statistic.
Usually we have no prior knowledge of space-time clusters location and then the test developed can not be applied since we have no cluster candidate C to use. Hence, our proposed test is based on the scan statistic

$$U = \sup_C \{U_C\} \tag{4}$$

which searches over all possible cylinders C (Kulldorff, 1997). In practice, the scanning in (4) is undertaken over a smaller class of cylinders for several reasons explained elsewhere.
The sampling distribution of U defined in (4) is intractable. As a consequence, its null hypothesis distribution is obtained by a Monte Carlo procedure conditionally on the realizations of the process spatial and temporal components. Under the null hypothesis, the sampling distribution of U is the distribution induced by random permutation of the times $t_i, i = 1, ..., n$ keeping fixed the spatial locations $(x_i, y_i), i = 1, ..., n$. The observed value u_1 of U is ranked amongst values $u_2, ..., u_B$ generated by recomputing the U statistic after B_1 independent random permutations of the times $t_i, i = 1, ..., n$. If u_1 ranks k-th largest, the one-sided exact attained significance level is k/m. This Monte Carlo method is computer intensive and naive algorithms should not be used for large data sets.

2 Application

For illustration, we use the crime incidence data from a large Brazilian city, Belo Horizonte, during 1995-2001 collected by the Poliícia Militar de Minas Gerais based on their police records of crime events. Each crime event was georeferenced by the coordinates of its occurrence place (em meters) and occurrence day. Four different data sets are used, investigating the space-time distribution of homicides as well as robberies of bakeries, drug stores
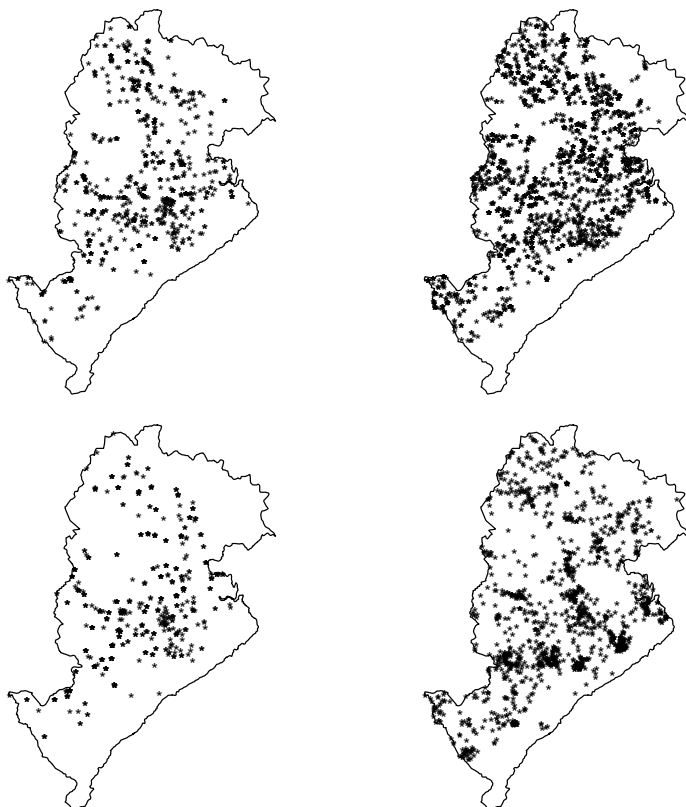
FIGURE 1. Maps of Belo Horizonte with four types of crime. The upper row shows the 765 drugstore robberies (left) and the 2216 bakery robberies (right). The bottom row shows the 582 lottery house robberies (left) and the homicides (right). The first three range from 1998 to 2000 while homicides data range from 1998 to 2001.

and lottery houses. Figure 1 shows the maps of all events for each one of the crimes.

Table 1 presents the results for the Knox test. It also shows results for our scan our scan procedure with a minimum of 5 events in each cylinder. We found C_1^* as a significant (at 0.05 level) space-time cluster in all four crimes, with bakery robberies presenting also C_2^* as a significant cluster (see Table 1). The number of events in the most significant cluster was 5, 7, 6, and 5 events for bakery, drugstore, lottery robberies, and homicide, respectively. The second significant cluster of bakery robberies had 5 events. Although the homicide space-time cluster presented borderline significance, we can

TABLE 1. Table with the p-values of Knox and scan tests. The results are separated according to either the thresholds used in the test (Knox) or the first (C_1^*) and second (C_2^*) most significant cylinders (scan test). The null hypothesis distribution was determined by 999 Monte Carlo permutations of the observed times t_i.

Crime	2 km, 20 days	3 km, 30 days	C_1^*	C_2^*
Bakery robbery	0.01	0.01	0.030	0.032
Drugstore robbery	0.01	0.01	0.012	0.154
Lottery robbery	0.05	0.22	0.028	0.220
Homicide	0.10	0.11	0.048	0.344

see that the scan test identified clusters in homicide and lottery robberies, whereas Knox test did not. This suggests that our method could be more sensitive to the presence of localized clusters than Knox test.

Concerning time, the shortest bursts of spatially localized violence was that associated with the two clusters of bakery robberies. They first and second clusters C_1^* and C_2^* lasted 8 and 17 days starting on February, 28 2000 and March 29, 2000 respectively. Drugstore and lottery robberies had longer clusters lasting 68 and 81 days starting on April 03, 1997 and May 23, 1995, respectively. The homicide cluster was detected on February 03, 2000, lasting 58 days.

The significant clusters of bakery robberies showed extreme patterns. Cluster C_1^* lasted only 8 days and, although occurring in different parts of the city, the second cluster started only 3 weeks after the first one had disappeared. This lasted only 17 days and contained five events related with 3 different stores, one of them being robbed three times during this period and five times during the total study period. The time lags between the five successive events in this second space-time cluster were 5, 2, 6, and 4 days.

References

Duczmal LH, Assunção RM (2003) A simulated annealing strategy for the detection of arbitrary shaped spatial clusters. *Computational Statistics and Data Analysis*, in press.

Kulldorff M (1997) A spatial scan statistic, *Communications in Statistics - Theory and Methods*, 26 (6): 1481-1496.

Mantel N (1967) The detection of disease clustering and the generalized regression approach. *Cancer Research*, 27, 209-220.

Parametric and Semi-Parametric approaches in the analysis of short-term effects of air pollution on health

Marc Saez[1], Michela Baccini[2] Annibale Biggeri[2] and Aitana Lertxundi[1]

[1] Research Group on Statistics, Applied Economics and Health (GRECS), University of Girona, Spain
[2] Department of Statistics "G. Parenti", University of Florence, Italy

Abstract: The standard approach in the analysis of short-term effects of air pollution on health is based on Generalized Additive Models (GAM), where seasonality and possibly other unobserved confounders are non-parametrically modeled. The aim of this paper is to compare, by a simulation study, performances of semi-parametric (GAM with penalized regression spline) and parametric approach (GAM with parametric regression spline) in term of estimation of air pollutant effect. We found that using semi-parametric approach can bring to biased estimates, unless a certain amount of undersmoothing is introduced. On the contrary negligible bias was found under the parametric approach, which appeared also robust to model misspecification.

1 Introduction

Currently GAMs have became a standard in the analysis of short-term effects of air pollution on health. In such models non-parametric functions of time (either cubic smoothing splines or locally weighted regression smoothers) are used to control for those unobserved confounders that could have a systematic temporal behavior. Recently critical points in using commercial statistical software which implements backfitting algorithm for fitting GAM were stressed (Dominici *et al.*, 2002; Ramsay *et al.*, 2003), encouraging use of alternative modeling strategies. They are based on simpler and more standard estimation methods, such as the fully parametric approach based on specification of Generalized Additive Models with regression splines (GAM+RS), or require much less computation for standard error estimation, such as the semi-parametric approach based on specification of GAMs with penalized regression splines (GAM+PRS). The objective of the paper is to compare, by means of a simulation study, the

performances of GAM+RS and GAM+PRS in estimating the parametric term which models the air pollutant effect in epidemiological time series regression.,

2   Methods

First we created pseudo data using the daily number of hospital admissions for respiratory diseases and the mean daily concentration of NO_2 from Barcelona (1995-1999). Adapting on real hospital admissions data a GAM with a penalized regression spline for time trend with df_0 degrees of freedom, we created a pseudo curve for seasonality, $f_0(t)$. A pseudo air pollution time series X_t was builded from the real NO_2 data, such that predefined amount of concurvity in data was obtained.

Then we generated 3000 outcome time series (Y_t) sampling from the following model:

$$Y_t \sim Po(\mu_{0t})$$

$$log(\mu_{0t}) = \alpha_0 + f_0(t) + \beta_0 X_t,$$

where β_0 denotes the "true" effect of air pollutant in term of log rate ratio. We analyzed each simulated data set using three different models: a GAM with a penalized regression spline for time trend with df_0 degrees of freedom, a GAM with a cubic regression spline with df_0 degrees of freedom and a GAM with a penalized regression spline whose degrees of freedom were selected by GCV. The first two models correspond to the situation in which the number of degrees of freedom to be assigned to the spline (df_0) is known. The following different scenarios were considered:

1. $\beta_0 = 0.0006$, concurvity $= 0.45$, $df_0 = 3, 4, 5, 7, 9$ per year

2. $\beta_0 = 0.0006$, $df_0 = 5$ per year, concurvity $= 0, 0.45, 0.7, 0.9$

3. $df_0 = 5$ per year, concurvity $= 0.45, \beta_0 = 0.0001, 0.0006, 0.006$

Finally, in order to assess robustness of parametric and semi-parametric approach to misspecification of degrees of freedom for the spline, we fitted on each simulated data set models with $df = 3, 4, \ldots, 15$ degrees of freedom per year (this analysis was performed under the *reference* scenario: $\beta_0 = 0.0006$, $df_0 = 5$ per year, concurvity $= 0.45$).

All the analyses were performed using the *mgcv* library implemented for R software by Wood (2000).

3   Results

When the number of degrees of freedom used for fitting data was correctly specified (Tab.1), the estimator of β in the semi-parametric model resulted

strongly biased for high amounts of smoothing. Similar results were obtained increasing concurvity amount in pseudo data (Tab. 2) and reducing the size of the air pollutant effect (Tab. 3). On the contrary, under the parametric approach negligible bias and good coverage of confidence intervals were found. Performances of GAM+PRS with smoothing parameter selected by GCV were comparable to the performances of the correctly specified GAM+RS.

TABLE 1. Results of simulation analysis varying the number of degrees of freedom in generating pseudo seasonality curve ($\beta=0.0006$; concurvity=0.45; $df_0=3,5,9$ per year).

df_0	% Relative Bias	Variance of Estimate	MSE (10^8)	Real Coverage of 95 % CI
GAM with natural cubic spline				
3	2.66	6.99	7.0	95.1
5	0.93	4.88	4.88	95.63
9	4.11	3.25	3.31	94.93
GAM with penalized regression spline				
3	155.57	6.63	93.7	5.86
5	15.53	4.68	5.56	93.43
9	5.01	3.19	3.28	94.8
GAM with penalized regression spline + GCV				
3	16.57	7.07	8.0	93.5
5	4.06	4.82	4.88	95.53
9	3.11	3.25	3.28	94.83

In the more realistic situation in which the actual number of degrees of freedom is unknown, the parametric approach appeared robust to mistakes in specifying the number of knots for the spline. On the contrary, the semiparametric approach produced biased estimates of air pollutant effect and bad confidence intervals if a small number of degrees of freedom was used to model seasonality (Fig. 1).

4   Discussion

Even if the number of degrees of freedom is correctly specified, the semiparametric approach can bring to strongly biased estimates and inappropriate confidence intervals for the parametric coefficient β. On the contrary, the estimator of air pollutant effect under the parametric model is negligibly biased (except that for unrealistically high concurvity) and the coverage of the 95% confidence intervals for β is always close to the real one. GAM+RS retains good property also under misspecification of the number of degrees of freedom for the regression spline, as shown by the robustness analysis.

TABLE 2. Results of simulation analysis varying concurvity amount in data (β=0.0006; concurvity=0.45,0.7,0.9; df_0=5 per year).

Concurvity	% Relative Bias	Variance of Estimate	MSE (10^8)	Real Coverage of 95 % CI
GAM with natural cubic spline				
0.45	0.93	4.88	4.89	95.63
0.7	-4.24	18.16	18.22	95.60
0.9	35.66	87.23	91.80	94.43
GAM with penalized regression spline				
0.45	15.53	4.68	5.55	93.43
0.7	81.76	16.59	40.66	81.07
0.9	292.82	57.13	365.80	45.00
GAM with penalized regression spline + GCV				
0.45	4.06	4.82	4.88	95.53
0.7	20.77	17.81	19.36	94.47
0.9	80.01	80.15	103.19	93.00

TABLE 3. Results of simulation analysis varying the air pollutant coefficient (β=0.0001,0.0006,0.006; concurvity=0.45; df_0=5 per year).

β	% Relative Bias	Variance of Estimate	MSE (10^8)	Real Coverage of 95 % CI
GAM with natural cubic spline				
0.0001	-5.51	5.07	5.08	95.47
0.0006	0.93	4.88	4.88	95.63
0.006	0.11	6.40	6.41	94.37
GAM with penalized regression spline				
0.0001	85.29	4.89	5.63	92.87
0.0006	15.53	4.68	5.55	93.43
0.006	1.12	6.31	6.76	92.90
GAM with penalized regression spline + GCV				
0.0001	14.04	5.04	5.06	95.17
0.0006	4.060	4.822	4.88	95.53
0.006	0.29	6.37	6.39	94.03

The semi-parametric approach works better for small values of the smoothing parameter used for generating pseudo seasonality curve. This outcome could indicate a certain tendency of semi-parametric approach to be more appropriate in presence of evident seasonality in data and/or reflect the beneficial effect of undersmoothing on the inference of the parametric component (Rice, 1986). This beneficial effect is emphasized by the improved
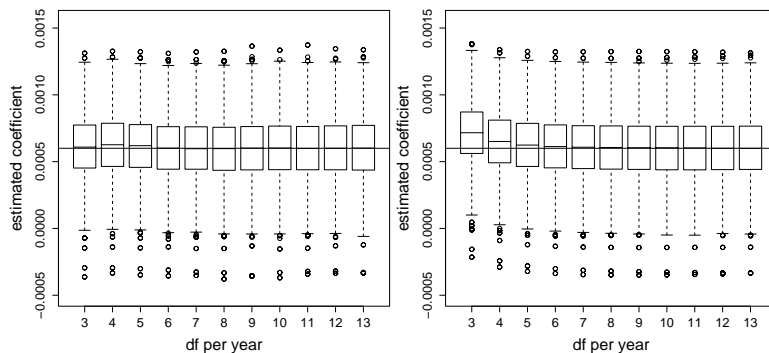
FIGURE 1. Distribution of the estimated effect of air pollution by number of degrees of freedom used for the smooth under GAM+RS (left) and GAM+PRS (right).

performance of GAM+PRS if combined with GCV, which is well-known to bring to undersmoothing.

In summary we can advance the following conclusions. Modeling seasonality by penalized regression splines or, plausibly, by other non-parametric functions, can provide biased estimates of air pollutant effect and misleading confidence intervals and should be avoided every time parametric alternatives are possible. The parametric approach is not affected by the same drawbacks as GAM+PRS and it is recommended. However, in presence of strong concurvity in data or very small effect size, sensitivity analysis changing number of knots is advisable.

References

Dominici, F., McDermott, A., Zeger, S.L. and Samet, J.(2002) On the use of generalized additive models in time-series studies of air pollution and health. *Am. J. Epidemiol.*, **156**, 193-203.

Ramsay, T.O., Burnett, R.T. and Krewski, D. (2003) The Effects of Concurvity in Generalized Additive Models Linking Mortality to Ambient Particulate Matter. *Epidemiology*, **14**, 18-23.

Rice, J. (1986) Convergence rates for partially splined models. *Statist. Probabil. Letters*, **4**, 203-208.

Wood, S.N. (2000) Modelling and Smoothing Parameter Estimation with Multiple Quadratic Penalties. *Journal of the Royal Statistical Society B*, **62**, 413-428.

Distributional results for FDR: application to genomic data

A. Bar-Hen[1], J.J. Daudin[1] and S. Robin[1]

[1] INA-PG/INRA Biométrie, 16 rue Claude Bernard, 75005 Paris, France

1 Introduction

Microarrays are part of a new class of biotechnologies that allow the monitoring of the expression level of thousands of genes simultaneously. Among the applications of microarrays, an important task is the identification of differentially expressed genes, i.e genes whose expressions is associated with the status of patients (treatment/control for example).

Multiple testing procedure is a classical problem for many high-dimensional data sets. The breakthrough of technology for image analysis or genomic have give a new interest for these questions. In this article we focus on differentially expressed genes but the proposed results are applicable for all multiple comparisons procedure.

The biological question of identification of differentially expressed genes can be restated as two-sample hypothesis testing procedure: does the gene is differentially expressed between the two situations. However, when thousands of genes in a microarray data set are evaluated simultaneously by fold changes and significance tests approach, multiple testing problems immediately arise and lead to many false positive genes. In this "one-by-one gene" the probability of detecting false positives rises sharply.

Basically, the various procedures proposed in the literature aim to test the null hypothesis

$$H_0(i) = \{\text{gene } i \text{ is not differentially expressed}\}.$$

These hypothesis is tested with two-sample tests. Corrections for heterogeneous variances, non-normality, non-independence of the tests were proposed (see S. Dudoit *et al.*, 2003 or Ge *et al.*, 2003 for recent review). Several solutions have been derived in the statistical literature to control the global type I error rate (see for example Holm, 1979 or, more recently, the false discovery rate (FDR, see Benjamini and Hochberg, 1995 or Tusher *et al.*, 2001).

FDR is defined as the fraction of false rejections among those hpotheses rejected. In the seminal paper (Benjamini and Hochberg, 1995) Benjamini

and Hochberg provided a distribtion free method for choosing a p value that guarantees that the FDR is less than a target level α. The same paper demonstrated that the B procedure is often more powerful than traditional methods that control familywise error (as Bonferroni method for example). Moreover, FDR is often of greater scientific relevance than the overall type I error rate. This work has been extended in various way. Benjamini and Yekutieli (2001) extendded the BH method to a class of dependent tests. Abaramovich, Benjamini, Donoho and johnstone (2000) established a connection between FDR and minimax point estimation. Efron, Tibshirani and Storey (2001) and Storey (2004) connected the FDR with bayesian quantities. Genovese and Wasserman (2001) showed that, asyptotically, the BH method corresponds to a fixed threshold method that rejects all p-values less than a given threshold and obtained some optimality results. In particular they proved that BH procedure is conservative. Since the aim of BH procedure is to control FDR, only a majorant can be found and the procedure is conservative. An alternative approach is to estimate the FDR and Storey (2002) and Storey, Taylor and Sigmund (2003) propose a family of point estimate, which is proved less conservative than BH procedure.

It is important to note that for both procedures, the idea is to derive results about the expected value of the proportion of false rejected hypothesis. It is an interesting result but not very useful for a particular experiment. In this talk we present results about the distribution, $f(.)$ of the proportion of false rejected hypothesis. Moments of $f(.)$ are easily derived. The expected value of $f(.)$ will be compared with classical procedure. From the second order moment we obtain confidence interval for the number of false rejected hypothesis.

Even if the procedure is stepwise, BH procedure and Storey procedure are based on distributional result for each step. In this presentation we obtain the joint distribution of all step and we obtain the conditional distribution of $f(.)$ at a given step conditionally on all previous step of the procedure. Simulations and example will be presented. In particular we compare our procedure to BH and Storey procedure for various cases.

References

Abramovich, F., Donoho, D., Johnstone, I. (2000). Adapting to unknown sparsity by controlling the false discovery rate. *Technical Report No. 2000-19, Dept. of Statistics, Stanford University.*

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JRSSB.* **57 (1)** 289–300.

Benjamini, Y. and Yekutieli, D. (2001) The control of the False Discovery Rate inMultiple Testing under Dependency, *Annals of Statistics.* **29 (4)** 1165–1188.

S. Dudoit, J. P. Shaffer, J. C. Boldrick (2003) Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18:1, 71-103.

Efron B., Tibshirani R., Storey J.D., Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, **96** 1151-1160

Ge, Y., S. Dudoit and Speed,T. P. (2003). Resampling-based multiple testing for microarray data hypothesis *Test* **12:1** 1-44

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society, Series B* **64** 499-517.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6** 65–70.

Storey, J.D. (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, **64** 479-498.

Storey, J.D., Taylor, J.E., and Siegmund D. (2004) Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B*, **66** 187-205.

Tusher, V. G., Tibshirani, R. Chu, G. (2001).Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*. **98**, 5116–5121.

Pairwise Likelihood for Generalized Linear Models with Crossed Random Effects

Ruggero Bellio[1] and Cristiano Varin[2]

[1] Dept. of Statistics, University of Udine, Italy. `ruggero.bellio@dss.uniud.it`
[2] Dept. of Statistics, University of Padova, Italy. `sammy@stat.unipd.it`

Abstract: Parameter estimation in Generalized Linear Models with crossed random effects is made difficult by the high-dimensional integrals required to obtain the full distribution of the response. We propose inference based on the pairwise likelihood, which only requires the computation of bivariate distributions. The estimators based on the pairwise likelihood are generally consistent, and the efficiency loss with respect to maximum likelihood estimation is usually not substantial. The method is applied to the famous salamander mating data.

Keywords: Binary data; GLMs; Pairwise likelihood; Crossed random effects.

1 Introduction

Generalized Linear Mixed Models (GLMMs) are widely used to accomodate overdispersion and correlation in data. These models are generated by adding random effects to the linear predictor of the corresponding Generalized Linear Model. A recent survey is given in McCulloch and Searle (2001).

For several years, computational aspects have represented a major obstacle in inference about GLMMs, in particular for the case of crossed random effects. Several methods have been proposed to overcome the numerical difficulties posed by high-dimensional integration. The popular PQL-type methods (McCulloch and Searle, 2001, §8.6) do not provide generally consistent estimation. Simulation-based algorithms for frequentist and Bayesian inference have been developed (e.g. Booth and Hobert, 1999, McCulloch and Searle, 2001, §10). However, they are quite computer intensive, so do not seem ready for daily use by practitioners, who often need to quickly estimate and analyze several different models at the model-building stage. This is particularly relevant with large sets of data.

In this work, we consider a composite likelihood approach based on marginal events (see Cox and Reid, 2003). Estimators based on suitable composite likelihood are generally consistent, and the efficiency loss with respect to maximum likelihood estimation is usually not substantial. The composite likelihood based on pairs of observations is denoted *pairwise likelihood* (Nott and Rydén, 1999). It has been successfully exploited by Renard et

al. (2004) for analyzing nested binary data through a GLMM with probit link. Here, we extend this approach to crossed random effects and general link functions for discrete data.

2   Pairwise likelihood inference

Let y_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m$ be a set of observed discrete data. Given a set of covariates $\{x_{ij}\}_{i,j}$, we assume a GLMM with conditional mean

$$g\{E(Y_{ij}|u_i, v_j)\} = x_{ij}^T \beta + u_i + v_j, \, i = 1, \ldots, n, \, j = 1, \ldots, m, \quad (1)$$

where $g(\cdot)$ is a suitable link function, while $u_i \sim N(0, \sigma_u^2)$ and $v_j \sim N(0, \sigma_v^2)$ are two sets of i.i.d Gaussian (crossed) random effects.
The likelihood function requires to compute an $n \times m$ intractable integral

$$L(\theta; y) = \int_{R^{n \times m}} \prod_{i=1}^{n} \prod_{j=1}^{m} p(y_{ij}|u_i, v_j; \beta) \phi(u_i; \sigma_u^2) \phi(v_j; \sigma_v^2) \mathrm{d}v_j \mathrm{d}u_i, \quad (2)$$

where $\theta = (\beta, \sigma_u, \sigma_v)$ and $\phi(\cdot; \sigma^2)$ represents the density function of a $N(0, \sigma^2)$ random variable. The computation of the above integral is challenging, since its dimension increases with the number of levels of the random factors. For this reason, we propose to use the pairwise likelihood, which is given by the product of the bivariate probabilities for all the possible pairs sharing at least one common random term

$$L_2(\theta; y) = \prod_{i=1}^{n} \prod_{j<j'}^{m} p(y_{ij}, y_{ij'}; \theta) \prod_{i<i'}^{n} \prod_{j=1}^{m} p(y_{ij}, y_{i'j}; \theta). \quad (3)$$

Each of the $n \binom{m}{2} + m \binom{n}{2}$ terms involved in $L_2(\theta; y)$ consists in a three-dimensional integral of the form

$$p(y_{ij}, y_{ij'}; \theta) = \quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad (4)$$
$$\int_{R^3} p(y_{ij}|u_i, v_j; \beta) \, p(y_{ij'}|u_i, v_{j'}; \beta) \phi(u_i; \sigma_u) \phi(v_j; \sigma_v) \phi(v_{j'}; \sigma_v) du_i dv_j dv_{j'}.$$

The computational effort required by the pairwise likelihood is much lower if compared to the "full" likelihood (2). In order to efficiently approximate the low-dimensional integrals forming $L_2(\theta; y)$, one can consider some standard deterministic quadrature rules, like Gauss-Hermite or the adaptive quadrature; see Evans and Swartz (2000). Morever, in the case of binary data and logit link (as for the Salamander mating data discussed in the next section), such integrals can also be approximated with high accuracy by normal scale mixtures; see Monahan and Stefanski (1992) and Drum and McCullagh (1993). Hereafter, we summarize the algorithm for obtaining the pairwise log-likelihood for θ.

Algorithm for computing the pairwise likelihood

1. Consider the random effect u_i, $i = 1, \ldots, n$.

   (a) Find all the pairs of observations sharing the random effect u_i.

   (b) For each pair $\{(i, j), (i, j')\}$, $j < j' = 1, \ldots, m$, evaluate the probability $p(y_{ij}, y_{ij'}; \theta)$.

   (c) Let $S_u(\theta; Y) = \sum_i^n \sum_{j<j'}^m \log p_{i,jj'}(\theta)$.

2. With similar steps as at point 1, obtain the quantity $S_v(\theta; Y)$ for the random effects v_j, $j = 1, \ldots, m$.

3. The log-pairwise likelihood is $\ell_2(\theta; Y) = S_u(\theta; Y) + S_v(\theta; Y)$.

From estimating equations theory, it follows that the *Maximum Pairwise Likelihood Estimator* (MPL) is consistent and asymptotically normally distributed. Denoting by ∇ the gradient operator, the variance matrix of the asymptotic distribution is given by $\mathrm{Var}(\theta) = H(\theta)^{-1} J(\theta) H(\theta)^{-1}$, where $H(\theta) = \mathrm{E}\{-\nabla^2 \log L_2(\theta; Y)\}$ and $J(\theta) = \mathrm{Var}\{\nabla \log L_2(\theta; Y)\}$. See Cox and Reid (2003) and reference therein for more details.

3   Example: salamander mating data

The salamander mating dataset has been already analysed by several authors, we refer to McCullagh and Nelder (1989, §14.5) for details on the experiment. The data consist in a collection of binary outcomes on the mating success between males and females from two populations of salamanders. A plausible model for this famous data is a GLMs with Bernoullian conditional density and two crossed effects, accounting for the male the female effect. The four fixed effects $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ included in the model are determined by the salamanders' gender and population; see Lin and Breslow (1996).

Following the same authors, we analyse the pooled dataset, treating all the three experiments done as they were obtained from different animals. In Table 1 we compare our MPL estimates with alternative methods, as reported by Lin and Breslow (1996) and Booth and Hobert (1999).

We found that the maximum pairwise likelihood estimates are close to those obtained with the other methods, with the exception of PQL, known to work poorly with binary data.

In order to compare the different methods, we also conducted a small simulation study following Lin and Breslow (1996) and Jiang (1998). We consider the same sample size of the pooled data. The mean values of the parameter estimates (with simulation standard errors in brackets) over $1,000$ replications are reported in Table 2. Here, MSM refers to the Method of

TABLE 1. Parameter estimates for the salamander mating data.

Estimate	α_0	α_1	α_2	α_3	σ_f^2	σ_m^2
Full Likelihood	1.03	−2.98	−0.71	3.65	1.40	1.25
Bayes/Gibbs	1.03	−3.01	−0.69	3.74	1.50	1.36
PQL	0.79	−2.29	−0.54	2.82	0.72	0.63
REML (D & M)	1.06	−3.05	−0.72	3.77	1.67	1.50
Pairwise Likelihood	1.07	−3.09	−0.73	3.81	1.69	1.58

TABLE 2. Results of a simulation study, 1,000 replications.

Parameter	α_0	α_1	α_2	α_3	σ_f^2	σ_m^2
True value	1.06	−3.05	−0.72	3.77	0.50	0.50
MSM (Jiang)	1.07	−3.13	−0.73	3.87	0.58	0.59
	(0.32)	(0.53)	(0.39)	(0.72)	(0.42)	(0.43)
PQL	0.94	−2.73	−0.64	3.38	0.33	0.32
	(0.27)	(0.40)	(0.34)	(0.49)	(0.22)	(0.22)
REML (D & M)	1.09	−3.14	−0.74	3.88	0.55	0.54
	(0.32)	(0.49)	(0.39)	(0.60)	(0.38)	(0.37)
Pairwise Likelihood	1.05	−3.07	−0.71	3.78	0.46	0.46
	(0.39)	(0.57)	(0.45)	(0.62)	(0.35)	(0.37)

Simulated Moments of Jiang (1998), and REML to the method of Drum and McCullagh (1993).

In this simulation study, we find a satisfactory performance for the MPL estimator, which seems slightly superior to the other methods under comparison.

4   Ongoing Research

We think that the pairwise likelihood is a promising method for inference in crossed random effect models. The advantages of this procedure are simplicity and computational efficiency. It follows that suitable bootstrap methods can be applied for improving inference; more details are given in Bellio and Varin (2003).

Ongoing research includes the development of model selection and model checking procedures based on the composite likelihood. Another interesting point to investigate is the application to large-scale problems.

References

Bellio, R. and Varin, C. (2003). A pairwise likelihood approach to generalized linear models with crossed random effects. Preprint 18.03, Department of Statistics, University of Udine. Submitted.

Booth, J.G. and Hobert, J.P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society series B*, **61**, 265-285.

Cox, D.R. and Reid, N. (2003). A note on pseudolikelihood constructed from marginal densities. Submitted manuscript. Available at `http://www.utstat.toronto.edu/reid/research.html`.

Drum, M.L. and McCullagh, P. (1993). REML estimation with exact covariance in the logistic mixed model. *Biometrics*, **49**, 677-689.

Evans, M. and Swartz, T. (2000). *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford: Oxford University Press.

Lin, X. and Breslow, N.E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, **91**, 1007-1016.

Jiang, J. (1998). Consistent estimators in generalized linear mixed models. *Journal of the American Statistical Association*, **93**, 720-729.

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

McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. New York: Wiley.

Monahan, J.F. and Stefanski, L.A. (1992). Normal scale mixture approximations to $F^*(z)$ and computation of the logistic-normal integral. In: *Handbook of the Logistic Distribution*, N. Balakrishnan (ed.), 529-540. New York: Marcel Dekker.

Nott, D.J. and Rydén, T. (1999). Pairwise likelihood methods for inference in image models. *Biometrika*, **86**, 661-676.

Renard, D., Molenberghs, G. and Geys, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics and Data Analysis*, **44**, 649-667.

Linking Gene-Expression Experiments with Survival-Time Data

Liat Ben-Tovim Jones[1], Shu-Kay Ng[1], Katrina Monico[1] and
Geoff McLachlan[1]

[1] Department of Mathematics and Institute for Molecular Bioscience, University
of Queensland, Brisbane 4072, Australia

Abstract: We apply a model-based clustering approach to classify tumour tissues
on the basis of microarray gene expression. The association between the clusters
so formed and patient survival (recurrence) times is examined. The approach is
illustrated using the lung cancer data set of Wigle et al. (2002). We show that the
prognosis clustering is a powerful predictor of the outcome of disease, in addition
to the stage of disease at presentation.

Keywords: Mixture models; EMMIX-GENE algorithm; Microarrays; Survival
analysis; Cox proportional hazards.

1 Introduction

In clinical medicine, accurately determining the stage of disease is crucial in
the management of cancer patients. Stage is defined using a combination of
clinical parameters (tumour size, lymph node involvement and the presence
of metastases). However, patients with the same stage of a particular can-
cer can have very different treatment responses and also clinical outcome.
There is much interest in determining whether microarrays can be used
as better indicators for outcome. Here we demonstrate how model-based
clustering in conjunction with survival analysis can be used to assess the
prognostic information in microarray data. We report in detail our results
for the lung cancer data set of Wigle et al. (2002). This data set formed
part of the CAMDA'03 challenge, and a fuller description of the methods
is given in Ben-Tovim Jones et al. (2004), and also their application to the
three other CAMDA'03 lung cancer data sets.

2 Cluster Analysis

Wigle et al. (2002) used cDNA microarrays to measure the gene expressions
for 39 tumour samples from patients diagnosed with various types of lung
cancer. We downloaded the data at `http://www.camda.duke.edu/camda03`,
and used the set of 2880 genes as in Wigle et al. (2002). For each patient, the

clinical outcome was given as the time between surgery and the recurrence. We label 1 to 24 the patients for which there has been a recurrence of the cancer, while those labelled 25-39 had no recurrence before the end of the study (their times to recurrence are censored). We input the data into the EMMIX-GENE algorithm of McLachlan et al. (2002). In the first screening step, 766 genes remained and these were then clustered into 20 groups. The means of these 20 groups (the metagenes) were used to cluster the tissues in the final step of EMMIX-GENE. Given the very small number of tumours (39) available here relative to the number of genes or indeed metagenes, some constraints had to be imposed on the component-covariance matrices in fitting a normal mixture model to cluster these tumours. We considered fitting to all 20 metagenes (a) mixtures of normals with equal component-covariance matrices; (b) mixtures of normals with (unrestricted) diagonal component-covariance matrices; and (c) mixtures of factor analyzers with equal component-covariance matrices for $q = 6$ factors. All three models led to two clusters, represented as

$$C_1 = \{15, 30 - 32, 34, 35, 37, 39\} \text{ and } C_2 = \{1 - 14, 16 - 29, 33, 36, 38\}.$$

Cluster C_1 corresponds to the good-prognosis group with 7 patients who are recurrence-free plus 1 patient who had experienced relapse of the tumour. This patient, however, was still alive at the end of the follow-up period. Cluster C_2 corresponds to the poor-prognosis group as it contains 23 of the 24 patients with recurrence, plus 8 patients with censored recurrence times. To further show that the first cluster C_1 corresponds to a recurrence-free group, we considered the long-term survival model

$$S(t) = \pi_1 + \pi_2 S_2(t), \tag{1}$$

where t is the time to recurrence, $S_2(t)$ is the conditional survival function for time to recurrence given recurrence will occur, and $\pi_2 = 1 - \pi_1$ is the probability of a recurrence. Under (1), a proportion π_1 of the patients will not have a recurrence; that is, their recurrence time is at infinity. The survival function $S_2(t)$ is taken to have the Weibull form,

$$S_2(t) = \lambda t^{\alpha-1} \exp(-\lambda t^\alpha). \tag{2}$$

The exact recurrence and survival times of two patients in C_2 were unknown and so they were excluded from all the survival analyses, leaving 37 patients with 15 of these censored. In Figure 1, we plot the fitted Weibull-based long-term survival model $\hat{S}(t)$ along with the Kaplan-Meier estimate. This shows excellent agreement between the nonparametric estimate as given by the Kaplan-Meier estimate and the parametric estimate $\hat{S}(t)$. In particular, from the asymptote of the curves, the probability π_1 of a patient being recurrence-free is approximately 0.2. Thus on average, one would expect to have approximately 8 recurrence-free patients in a set of 39. Here the cluster C_1, which is conjectured as corresponding to the
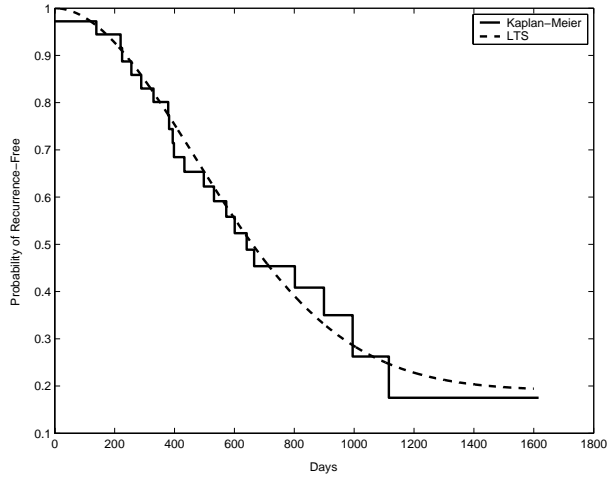
FIGURE 1. Fitted LTS model versus Kaplan-Meier.
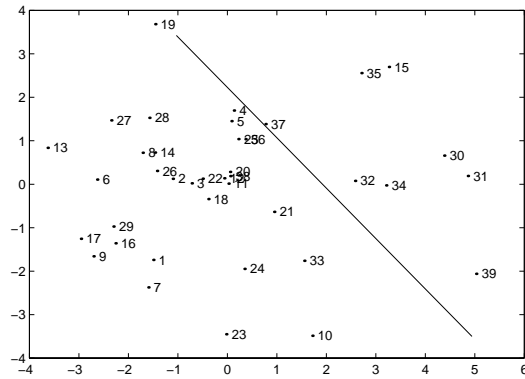
FIGURE 2. PCA of tissues based on 20 metagenes.
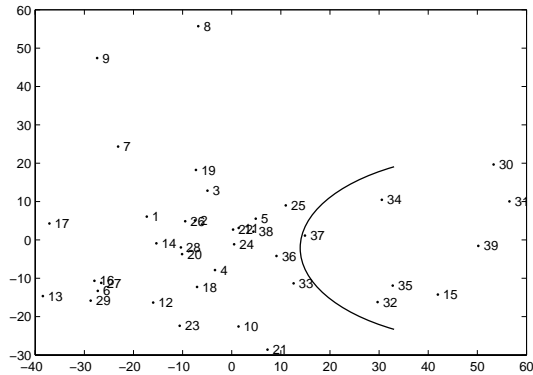
FIGURE 3. PCA of tissues based on all genes (via SVD).

recurrence-free group, has indeed 8 members in it. Interestingly, 5 of the censored patients clustered into C_2 were also put together in a cluster corresponding to early recurrence in the hierarchical clustering of Wigle et al. (2002). This long-term survival model (1) can be used also to estimate the posterior probability that a patient with a censored recurrence time will be recurrence-free. Unfortunately, unless the censored time is very long, these estimated posterior probabilities are equal, being around 0.5. Patient (P81 AC) who has a censored time of 1,161 days has a high posterior probability of being recurrent-free so her membership of cluster C_1 would appear to be atypical. To further investigate the validity of our clustering of the 39 tumours, we considered a plot of the first two principal components (PCs) of the tumours obtained by a singular-value decomposition based on (a) the 20 metagenes and (b) all the genes, as given in Figures 2 and 3, respectively. In each of these two figures, we have imposed the allocation boundary that will give the clustering that we have obtained above. In each case, it can be seen that this boundary represents a reasonable partition of the data into two clusters in the space of the first two PCs.

3   Survival Analysis

For the 37 patients with survival data available, we clustered 29 as poor prognosis (C_2) and 8 as good prognosis (C_1). We use the Kaplan-Meier estimate to provide an estimate of the overall probability of being recurrence-free following surgery. Given that there is only one recurrence in C_1, it should have a significantly better Kaplan-Meier estimate than C_2, and this is confirmed in Table 1. These two Kaplan-Meier estimates are plotted in Figure 4. The Kaplan-Meier curves were compared with the use of the log-rank test.

TABLE 1. Non-parametric Survival Analysis

Cluster	No. of Patients (Censored)	Mean Time to Recurrence ($\pm$ SE)
C_1	8 (7)	1388 $\pm$ 155.7
C_2	29 (8)	665 $\pm$ 85.9

We also fitted the proportional hazards model of Cox (1972), using covariates to represent the clinical data and a zero-one indicator variable to membership of cluster C_1 or not. The fit for the final form of this model is given in Table 2. The significance of estimated hazard ratios were tested using the Wald test. All calculations in the survival analysis were performed with the S Plus statistical package. It can be seen that membership of cluster C_1 (the poor-prognosis cluster) was the only significant factor affecting the event of being recurrence-free ($P = 0.06$).
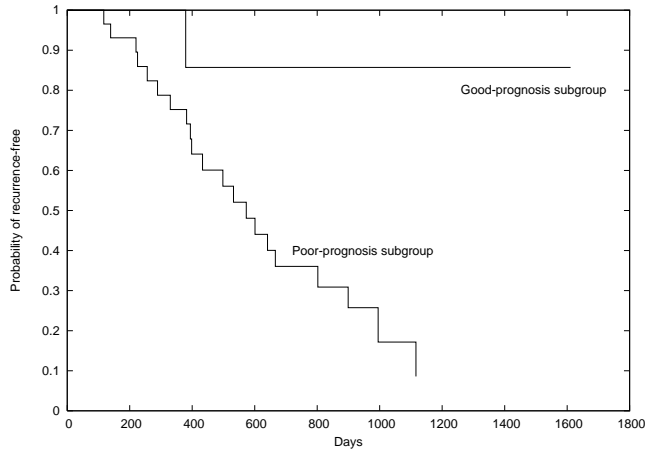
FIGURE 4. Kaplan-Meier curves of recurrence-free for the two clusters.

TABLE 2. Multivariate Cox Hazards Analysis of the Risk of Recurrence

Variable	Hazard Ratio (95%CI)	P-Value
Poor (vs. good prognosis cluster)	6.8 (0.9-51.8)	0.06
Stages 2 or 3 (vs. Stage 1)	1.1 (0.4-2.7)	0.88

4   Conclusions

We were able to use a model-based clustering approach to to identify patient clusters with clinical outcomes of recurrence versus non-recurrence of tumour. The gene-expression data provided prognostic information, beyond the clinical indicator of stage. A limiting factor in the analyses was the small numbers of tumours available. Further, the high proportion of censored observations limited the comparison of survival rates.

References

Ben-Tovim Jones, L., Ng, S.K., Ambroise, C., Monico, K., Khan, N., et al. (2004). Use of microarray data via model-based classification in the study and prediction of survival from lung cancer. In: *Methods of Microarray Data Analysis IV*, K.F Johnson and S.M. Lin (Eds.). Dordrecht. Kluwer. To appear.

McLachlan, G.J., Bean, R.W., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413–422.

Wigle, D.A., Jurisica, I., Radulovich, N., Pintilie, M., et al. (2002). Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Research*, **62**, 3005–3008.

Rank tests of conditional independence for continuous variables

Wicher P. Bergsma[1]

[1] EURANDOM, PO Box 513, 5600 MB Eindhoven, bergsma@eurandom.tue.nl

Abstract: Many rank tests are available for the testing of unconditional independence, for example tests based on Kendall's tau or Spearman's rho, but for conditional independence this is unfortunately not the case. This paper introduces a general method based on estimation of the conditional distribution functions of response variables given control which allows arbitrary rank tests of unconditional independence to be applied to the testing of conditional independence.

Keywords: copula, kernel estimation, Kendall's tau, Spearman's rho

1   Introduction

For a given triple of random variables (X, Y, Z) we consider the problem of testing the hypothesis of conditional independence of Y and Z controlling for X based on n independent and identically distributed (iid) data points $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)$. Following Dawid (1979), this hypothesis is denoted as

$$Y \perp\!\!\!\perp Z | X$$

Even though a wide array of tests is available for the testing of independence between two random variables (see, for example, Kendall and Gibbons, 1990, Nelsen, 1998 [Chapter 6] or Schweitzer and Wolff, 1981), the choices are much more limited for the testing of conditional independence.

In particular, for continuous variables and without strong distributional assumptions, there appears to be only one choice, namely the test based on the partial correlation coefficient; with marginal regressions $Y = g(X) + \epsilon_1$ and $Z = h(X) + \epsilon_2$, it is defined as

$$\rho_{23|1} = \frac{\mathrm{cov}(\epsilon_1, \epsilon_2)}{\sqrt{\mathrm{var}(\epsilon_1)\mathrm{var}(\epsilon_2)}} \tag{1}$$

Evaluation of the test requires the estimation of the regression curves, which has to be done non- or semi-parametrically unless a specific parametric form is known to hold a priori.

Another test statistic for conditional independence, based on Kendall's tau was proposed by Goodman (1959) and further discussed by Goodman and

Grunfeld (1961). However, the distributional assumptions underlying this test appear somewhat complex (Gripenberg, 1992).

In this paper we propose a new method to obtain more general tests of conditional independence than those based on (1). The theoretical background is given in Section 2. A practical procedure, based on a simple kernel estimation method, is given in Section 3. The estimation problem is (distantly) related to median regression.

2   The partial copula

For the conditional distribution functions of Y and Z we write

$$F_{2|1}(y|x) = \Pr(Y \leq y | X = x)$$
$$F_{3|1}(z|x) = \Pr(Z \leq z | X = x)$$

A basic property of $U = F_{2|1}(Y|X)$ and $V = F_{3|1}(Z|X)$ is given in the following lemma.

Lemma 1 *Suppose, for all x, $F_{2|1}(y|x)$ is continuous in y and $F_{3|1}(z|x)$ is continuous in z. Then U and V have uniform marginal distributions.*

The importance of the introduction of U and V lies in the following theorem.

Theorem 1 *Suppose, for all x, $F_{2|1}(y|x)$ is continuous in y and $F_{3|1}(z|x)$ is continuous in z. Then $Y \perp\!\!\!\perp Z | X$ implies $U \perp\!\!\!\perp V$.*

The proof is given below. Theorem 1 implies that a test of unconditional independence of U and V is a test of conditional independence of Y and Z given X. A test of independence of U and V can be done by any standard procedure.

For continuous random variables Y and Z with marginal distribution functions F_2 and F_3, the *copula* of their joint distribution is defined as the joint distribution of $F_2(Y)$ and $F_3(Z)$. The copula is said to contain the grade (or rank) association between Y and Z (for an overview, see Nelsen, 1998). For example, Kendall's tau and Spearman's rho are functions of the copula. The following definition gives an extension of the copula concept.

Definition 1 *The joint distribution of U and V is called the* partial copula *of the distribution of Y and Z given X.*

Hence, the partial copula is an appropriate basis for studying conditional dependence.

It should be noted that since $U \perp\!\!\!\perp V$ does not imply $Y \perp\!\!\!\perp Z | X$, a test of the hypothesis $U \perp\!\!\!\perp V$ cannot have power against all alternatives of the hypothesis $Y \perp\!\!\!\perp Z | X$. In particular, this is so for alternatives with interaction, that is, where the association between Y and Z depends on the

value of X. We should expect most power against alternatives with a constant conditional copula, i.e., alternatives for which the joint distribution of $(F_{2|1}(Y|x), F_{3|1}(Z|x))$ does not depend on x.

Proof of Lemma 1: By continuity of $F_{2|1}(y|x)$ in y, and with F_1 the marginal distribution function of X,

$$
\begin{aligned}
\Pr(U \le u) &= \Pr(F_{2|1}(Y|X) \le u) = \int \Pr(F_{2|1}(Y|x) \le u) dF_1(x) \\
&= \int u \, dF_1(x) = u
\end{aligned}
$$

i.e., the marginal distribution of U is uniform. The uniformity of the distribution of V is shown analogously.

Proof of Theorem 1: By Lemma 1, U and V are uniformly distributed. Hence if $Y \perp\!\!\!\perp Z | X$ the joint distribution function of U and V simplifies as follows:

$$
\begin{aligned}
\Pr(U \le u, V \le v) &= \Pr(F_{2|1}(Y|X) \le u, F_{3|1}(Z|X) \le v) \\
&= \int \Pr(F_{2|1}(Y|x) \le u, F_{3|1}(Z|X) \le v) dF_1(x) \\
&= \int \Pr(F_{2|1}(Y|x) \le u) \Pr(F_{3|1}(Z|X) \le v) dF_1(x) \\
&= \int uv \, dF_1(x) = uv = \Pr(U \le u) \Pr(V \le v)
\end{aligned}
$$

This completes the proof.

3   Kernel estimation of the conditional distributions

In general, a rank test for independence between Y and Z is a function of the copula and is based on the rank transformations $F_2(Y)$ and $F_3(Z)$, where F_2 and F_3 are the marginal distribution functions of Y and Z, respectively. A broad class of rank test of unconditional independence can be written as a U-statistic of degree r, in particular, for an appropriate function ϕ, in the form

$$
T = \binom{n}{r}^{-1} \sum \phi[(F_2(Y_{i_1}), F_3(Z_{i_1})) \ldots, (F_2(Y_{i_r}), F_3(Z_{i_r}))] \tag{2}
$$

where the summation is over all subsets $\{i_1, \ldots, i_r\}$ of $\{1, \ldots, n\}$. For example, Spearman's rho is written as

$$
\rho_S = \binom{n}{2}^{-1} \sum_{i \ne j} (F_2(Y_i) - F_2(Y_j))(F_3(Z_i) - F_3(Z_j))
$$

and Kendall's tau can be written as

$$\tau = \left(\begin{array}{c} n \\ 2 \end{array} \right)^{-1} \sum_{i \neq j} \text{sign}(F_2(Y_i) - F_2(Y_j))(F_3(Z_i) - F_3(Z_j))$$

Another important example is Hoeffding's coefficient of independence (see Manoukian, 1986). The unknown distribution functions are replaced by the empirical distribution functions.

Using the results of the previous section, a rank test of conditional independence has the form (2) with $F_2(Y_i)$ replaced by $F_{2|1}(Y_i|X_i)$ and $F_3(Z_i)$ replaced by $F_{3|1}(Z_i|X_i)$. However, the latter cannot be estimated by the empirical distribution functions, since (assuming continuity of X) for each X_i there is, with probability 1, only one observed pair (Y_i, Z_i). Instead, we propose the following kernel estimator:

$$\hat{F}_{2|1}(y|x) = \frac{\sum_{i=1}^{n} K[(\hat{F}_1(x) - \hat{F}_1(X_i))/h]I(Y_i < y)}{\sum_{i=1}^{n} K[(\hat{F}_1(x) - \hat{F}_1(X_i))/h]}$$

where $h > 0$ is the bandwidth, usually dependent on n, K is the kernel function, which can be a density symmetric around zero, I is the indicator function and

$$\hat{F}_1(x) = n^{-1} \sum_{i=1}^{n} I(X_i < x)$$

is the empirical distribution function of X. A suitable choice for K is often the standard normal distribution.

Note that the above problem is related to median regression; there, for all x a solution y is required of the equation

$$F_{2|1}(y|x) = \tfrac{1}{2}$$

Also note that the test based on T_1 is quite different from the test based on Kendall's partial tau (Kendall, 1942), which is not necessarily zero under conditional independence (Korn, 1984)

References

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society, Series B*, **41**, 1-31.

Goodman, L. A. (1959). Partial tests for partial tau. *Biometrika*, **46**, 425-432.

Goodman, L. A. and Grunfeld, Y. (1961). Some nonparametric tests for comovements between time series. *Journal of the American Statistical Association*, **56**, 11-26.

Gripenberg, G. (1992). Partial rank correlations. *Journal of the American Statistical Association*, **87**, 546-551.

Kendall, M. G. (1942). Partial rank correlation. *Biometrika*, **32**, 277-283.

Kendall, M. G. and Gibbons, J. D. (1990). *Rank correlation methods.* New York: Oxford University Press.

Korn, E. L. (1984). The ranges of limiting values of some partial correlations under conditional independence. *The American Statistician*, **38**, 61-62.

Manoukian, E. B. (1998). *Mathematical nonparametric statistics.* New York: Gordon and Breach Science Publishers.

Nelsen, R. B. (1998). *An introduction to copulas.* New York: Springer.

Schweizer, B. and Wolff, E. F. (1981). On nonparametric measures of dependence for random variables. *Annals of Statistics*, **9**, 879-885.

Investigating gene-specific variance via Bayesian hierarchical modelling

M. Blangiardo[1], A. Biggeri[1], S. Toti[1], C. Lagazio[2], B. Giusti[3]

[1] Department of Statistics "G.Parenti", University of Florence, Italy
[2] Department of Statistical Science, University of Udine, Italy
[3] Dipartimento Area Critica Medico Chirurgica, University of Florence, Italy

Abstract: We are using Bayesian hierarchical models to estimate gene-specific variance in calibration experiments, where two samples from the same population are labelled with the two Red and Green dyes. The estimates from these experiments are incorporated as prior knowledge in comparative ones. This procedure allows to use a different variance for each gene, and could be very useful with the aim of collecting some prior information about new experiments to be performed.

Keywords: Microarray, Bayesian Hierarchical models; Gene-specific variance.

1   Introduction

Microarray studies permit to quantify expression levels on a global scale by measuring transcript abundance of thousands of genes simultaneously. The description, classification and study of the relationships between genes are the new tasks made possible by innovative research tools. A difficulty when analyzing expression measures obtained by cDNA arrays is how to model the variance function for the whole set of genes. In such contexts, it is usually unrealistic to assume a common variance and would be better to consider different measure of variability for each gene. To this aim, Tseng et al. (2001) introduced a calibration experiment, in which the probes hybridized on the two channels come from the same population (self-self experiment). From such an experiment, it is possible to estimate the gene-specific variance, to be incorporated in comparative experiments on the same tissue, cellular line or species. We present a Bayesian hierarchical model to use the information on gene-specific variability from a calibration experiment to be incorporated as prior knowledge in comparative experiments. We apply the methodology to a real example and compare our results to those obtained with Tseng's approach.

2   Materials and methods

Mononuclear cells were obtained from peripheral blood of 10 healthy subjects by density gradient centrifugation on Ficoll-Hypaque. Cells from each

subjects were incubated in RPMI 1640 at 37 C in a humidified atmosphere with 5% CO2 for 3 hours in presence or absence of lipopolysaccharide (LPS, 10 mg/ml). Total RNA was extracted and equal amount of total RNA, from stimulated or unstimulated cells, from different subjects was pooled. Total RNAs were retro-transcribed with amino-allyl-dUTP, hydrolyzed, purified and labelled with NHS-Cyanine dyes (Cy3 and Cy5). Then, the two probes were purified, mixed and hybridized on the arrays. After incubation, arrays were scanned by the 4000B scanner (Axon). Image analysis was performed by GenePix 4.1 software. 5 arrays were printed. For calibration purposes 3 self-self arrays were performed using probes from cells incubated in absence of LPS. 2 arrays were fabricated for the comparison experiments, using dye-swap. All the 5 arrays were subjected to quality controls following the criteria suggested by Simon et al. (2003), to eliminate low-intensity genes. We did not expect to find any differentially expressed gene in calibration arrays.

The first stage in the analysis was to estimate gene-specific variance from the calibration experiments. To this purpose we specified a linear ANOVA model (Kerr et al. 2000, Lewin et al. 2003) where the unnormalized log gene expression intensity for each array

$$y_{gs} \quad \sim \quad N\left(\mu_{gs}, \sigma_g^2\right) \tag{1}$$

were modelled as Gaussian for gene g and channel $s = 1, 2$.

Moreover, specific terms in the linear predictor

$$\mu_{gs} \quad = \quad \alpha_{ag} + \delta_s + \nu_g \tag{2}$$

were introduced to mimic the normalization procedure, where α_{ag} was the gene-specific array effect and δ_s was the dye-effect; ν_g was the normalized gene effect.

The gene-specific variance was assumed to follow the Lognormal distribution $\sigma_g^2 \sim \log N(m, s^2)$ where $m \sim N(0, 10000)$ and $1/s^2 \sim G(0.001, 0.001)$ were noninformative hyperpriors, while $\nu_g \sim N(a, b^2)$, a was non informative Gaussian and $1/b^2$ was a non informative Gamma. Finally, all the other normalization parameters were modelled as non informative Gaussian distributions. We compared the performance of this model with that of a model specifying a common variance σ^2 for all genes. To compare models we used Deviance Information Criterion (Spiegelhalter et al. 2002).

In the second stage of the analysis we built up a hierarchical Bayesian model for the comparative experiment, incorporating posterior densities of m e s^2 from the calibration experiment. For this model we had informative hyperpriors and we included a treatment effect τ_g in the linear predictor:

$$\mu_{gs} = \alpha_{ag} + \tau_g + \delta_s + \nu_g. \tag{3}$$

Summaries from the posterior densities of τ_g can be used to identify differentially expressed genes.

We analyzed our data also using the Tseng's Bayesian hierarchical model. To perform the analysis we used WinBugs (see Spiegelhalter et al. 2003) and R (see `http://cran.r-project.org`).

3   Results

The analysis was performed on 2887 genes, which have not presented missing values in any of the 5 arrays and that passed quality controls. From the calibration experiment, we found gene specific variances ranging from 0.015 to 0.03 (figure 1 reports the distribution of the posterior gene-specific variances). The comparison to the common variance model was performed by Deviance Information Criterion (DIC) and showed a better behavior of the gene-specific variance model.

The analysis of the comparative experiment resulted in a list of 37 differentially expressed genes. The comparison to Tseng's model brought out some differences in terms of altered genes. In particular, the number of differentially expressed genes with the two methods is shown in table 1: 26 genes emerge as significative under both approaches. Literature confirmed an alteration in gene expression profile after LPS stimulation on peripheral blood mononuclear cells for 11 out of the 26 genes. As concerned the genes emerged as differentially expressed only using our Bayesian hierarchical model, data from the literature are available confirming the upregulation after LPS stimulation for 5 genes.

The differences are related to the genes with a low, positive or negative relative expression. Actually these genes are the most influenced by changing assumptions on gene variances.

Also for the comparative experiments, we have found a better behavior for our model with respect to the Tseng's one. In particular, the DIC statistics is 34560 for our model and reaches 35010 for the other.

4   Discussion

The observed differences in number of differentially expressed genes among the approaches are related to different variance modelling. Both Tseng's model and our Bayesian hierarchical model, consider a gene specific variance and seem to carry out sensitive estimates. However, the difference between our model and Tseng's one are related to the initial assumptions: for our model the likelihood formulated on the single channel intensity, while in the other model the likelihood is based on the normalized log ratio. The gene variance is also modelled differently: Tseng et al. consider the gene specific variance and the average variance from the calibration arrays as observed quantities; they compute a weighted average between these two components and incorporate it as data to estimate the prior distribution of the variance to be used as information for the comparative experiment. The

prior distribution of the variance is a transformed chi-squared. On the other side our Bayesian hierarchical model starts from the calibration experiment, computes a posterior distribution of variance parameters and incorporates that in the model for the comparative experiment as prior knowledge. Besides, our prior distribution of variance is lognormal. Finally our model is very general and easily allows us to perform sensitivity analysis, to change prior distributions or likelihood. Eventually, our approach seems useful to be followed when considering a sequence of experiments (e.g. time course experiments): it permits to update estimate of the variances and to take under control sources of variations that can be introduced between different experiments. In order to better evaluate the strength of the two different approaches, further information are needed about the genes emerged only in one model. Real time PCR experiments on these genes are in progress to confirm the results.

References

Kerr, M.K., Martin, M. and Churchill, G.A. (2000). Analysis of variance for gene expression microarray data, *Journal of Computational Biology*, **7**, 819-837.

Lewin, A., Richardson, S., Marshall, C., Glazier, A. and Aitman, T. (2003). Bayesian Modelling of Differential Gene Expression, *submitted*.

Simon, R.M., Korn, E.L. and McShane, L.M. (2003). *Design and Analysis of DNA Microarray Investigations*. Springer-Verlag.

Spiegelhalter, D., Best, N., Carlin, B. Van Der Linde, A. (2002). Bayesian measures of model complexity and fit., *Journal of the Royal Statistical Society B*, **64**, 1-34.

Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2003). *WinBUGS, version 1.4*. MRC Biostatistics Unit, Cambridge,UK.

Tseng, G.C., Oh, M.K., Rohlin, L., Liao, J.C. and Wong, W.H. (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects., *Nucleic Acids research*, **29**, 2549-2557.

TABLE 1. Number of differentially expressed genes (Comparison of two different approaches)

	Tseng *et al.*	Hierarchical Bayesian
Tseng *et al.*	41	
Hierarchical Bayesian	26	37
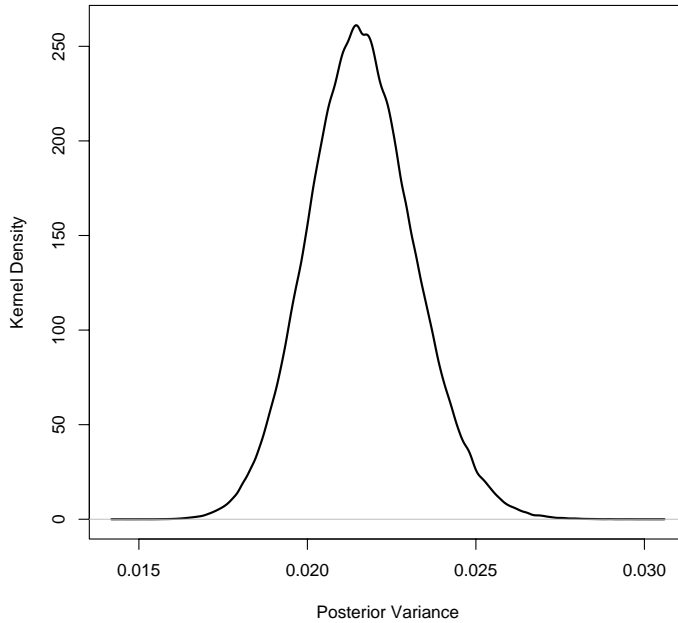
FIGURE 1. "Marginal" posterior variances.

On the clustering term in ecological analysis: how do different prior specifications affect results?

Dolores Catelan[1], Annibale Biggeri[1] and Corrado Lagazio[2]

[1] Dept. of Statistics "G. Parenti", University of Florence, Viale Morgagni, 59 - 50134 Firenze, Italy; email: catelan@ds.unifi.it
[2] Dept. of Statistics, University of Udine, Via Treppo, 18 - 33100 Udine, Italy; email: lagazio@dss.uniud.it

Abstract: In this work we give an example of how different prior assumptions on the clustering term of a hierarchical bayesian model with time dependent covariate could affect the results of the analysis.

Keywords: Ecological analysis, Hierarchical Bayesian model, Spatial random terms, Conditional autoregressive models, Time-dependent covariates.

1 Introduction

The aim of ecological studies is to describe the relationship between geographical variation of disease risk and concomitant variation in the level of exposure to a particular factor: for example, an environmental agent or a life-style related characteristic. In our analysis we use education as a proxy of socioeconomic factors.
Both disease rates and covariates could exhibit a strong spatial autocorrelation. If ignoring this aspect might produce incorrect inferences (see Clayton et al., 1993), care must be taken in modelling spatially structured overdispersion since the random term could absorb part of the association and bias the estimate of the effect of the exposure (see Wakefield, 2003).
In this work we give an example of how different prior assumptions on the clustering term of a hierarchical bayesian model with a time dependent covariate could affect the results of the ecological analysis.

2 Data

Lung cancer death certificates are considered for males resident in 287 municipalities of the Tuscany Region (Italy) from 1971 to 1999. Mortality data are aggregated in six calendar periods (1971-74, 1975-79,..., 1995-99). We use internal indirect standardization to calculate the expected number of cases.

For the aim of our analysis we have considered the proportion of population with primary school degree in the years 1951, 1961, 1971, 1981 and 1991 as the exposure variable. Since mortality and education are recorded in different time points we need to estimate a value of the education score for years $1956, 1966, 1976, 1986, 1996$ and for each municipality.

3   Space-time models with time-dependent covariates

We propose a generalization of the model of Knorr-Held (2000) in which we replace the space-time interaction with a time-dependent covariate, considered at different lags, to take into account for the latency between exposure and disease onset (for details see Dreassi et al. (2003)).

The model assumes that the number of observed cases in the i-th area $(i = 1, \ldots, 287)$ and j-th period $(j = 1971\text{-}74, 1975\text{-}79, 1980\text{-}84, 1985\text{-}89, 1990\text{-}94, 1995\text{-}99)$ $O_{i,j}$ follows a Poisson distribution with mean $E_{i,j}\theta_{i,j}$, where $E_{i,j}$ indicates the expected number of cases under indirect standardization and $\theta_{i,j}$ the relative risk. A random effects model is assumed for the logarithm of the relative risk

$$\log(\theta_{i,j}) = u_i + v_i + p_j + \beta_j \ \mathbf{x}'_{i,j-l} \ \lambda_j. \tag{1}$$

The parameters β_j define the relationship between mortality in the j-th period and education observed 0, 5, 10, 15 years before: we are taking into account that the process of carcinogenesis involves a latency time (e.g. a time equal to l), hence mortality on time j would result in association with a covariate observed at time $j - l$ $(l = 0, 5, 10, 15)$.

The prior on each coefficient β_j is a flat Normal distribution.

The *heterogeneity* term u_i represents an unstructured spatial variability component modelled as Normal (μ_u, δ_u) where δ_u is the precision parameter and is assumed to follow a flat Gamma distribution. The term p_j represents the effect of the j-th period which is assumed to follow a first order random walk with independent normal increments (see Knorr-Held, 2000 for further details). The vector $\mathbf{x}_{i,j-l} = (x_{i,j}, x_{i,j-5}, x_{i,j-10}, x_{i,j-15})$ contains the education scores for the i-th area observed at the four considered lags. Terms $\lambda_j \sim \text{multinomial}(\pi_j, 1)$ and $\pi_j = (\pi_{j0}, \pi_{j5}, \pi_{j10}, \pi_{j15})' \sim \text{Dirichlet}(1,1,1,1)$ represent respectively weights and probabilities for each lag and for each period whose estimation is one of the purpose of the analysis.

3.1   Prior distribution of the clustering component

Assume we have a set of area-specific spatially correlated Gaussian random effects v_i for $i = 1, \ldots, N$ (the $\mathbf{v}$ term is called *clustering* term). Suppose their joint distribution may be expressed as $\mathbf{v} \sim \text{MVN}(\mu, \delta_v \mathbf{\Sigma})$ where MVN stays for Multivariate Normal distribution, μ is the mean vector, $\delta_v > 0$

controls the overall variability of the v_i and $\mathbf{\Sigma}$ is an $N \times N$ positive definite matrix.

Let define the between area covariance matrix as $\delta_v \mathbf{\Sigma} = \delta_v (\mathbf{I} - \rho \mathbf{W})^{-1} \mathbf{M}$ where $\mathbf{I}$ is the identity matrix, $\mathbf{W}$ is a weight matrix with elements W_{ik} reflecting spatial association between areas i and k, $\mathbf{M}$ is a diagonal matrix with elements M_{ii} proportional to the conditional variance of $v_i | v_k$ and ρ controls overall strength of spatial dependence.

Different specifications are possible for $\mathbf{\Sigma}$. In particular, we may assume a parametric form for the elements of the matrix. In this case a common assumption is $\Sigma_{ik} = \exp[-(\phi \, d_{ik})^\nu])$ where d_{ik} is distance between the centroids of areas i and k, $\phi > 0$ controls the rate of decline of correlation with distance and $\nu \in (0, 2]$ controls the amount of spatial smoothing (see Journel et al. (1978)). We have fitted the ordinary exponential model ($\nu = 1$) with a Uniform prior distribution for ϕ.

Otherwise, following the *conditional* formulation, we do not need to specify the elements of the covariance matrix $\mathbf{\Sigma}$ but work just on $\mathbf{W}$, $\mathbf{M}$ and ρ. Besag et al. (1991) propose an Intrinsic CAR model (ICAR) for v_i in which $\mathbf{\Sigma}$ is not positive definite. This model corresponds to choose $W_{ik} = 1/n_i$ if $i \sim k$ ($i \sim k$ indicates that the i-th and k-th areas are adjacent) and 0 otherwise, $M_{ii} = 1/n_i$ and $\rho = 1$ and leads to a Normal $(\bar{v}_i, \delta_v n_i)$ conditional distribution for $v_i | v_k$, where $\bar{v}_i = \sum_{k \sim i} \frac{v_k}{n_i}$ is the mean of the terms of the adjacent areas and n_i is their number.

Alternative choices of $\mathbf{W}$ and $\mathbf{M}$ lead to a full-rank covariance matrix. Here we follow the assumption of Stern et al. (1999) defining $M_{ii} = 1/E_i$, $W_{ik} = (E_k/E_i)^{1/2}$; in this case we have also to specify a prior distribution for ρ, which we assume to be uniformly distributed in (ρ_{min}, ρ_{max}). E_i is the expected number of cases in the i-th area for the entire study period. To make comparisons between models we have made use of the Expected Predictive Deviance (EPD) (see Laud et al. (1995)).

4   Results

In Figure 1 we report the education score in 1951 (1961, 1971, 1981 and 1991 exhibit the same spatial structure) and the disease risks estimated with the standard model of Besag et al. (1991) without considering the covariate effect and collapsing the data over the entire period 1971-99.

Exposure and mortality show similar spatial patterns, with a higher level of risk in areas with a higher level of education: we then expect a positive association between the education score and disease risks.

Surprisingly, when we specify the ICAR prior ($\rho = 1$) the estimates of β parameters are negative (see Table 1). When we fit this model assuming different values of ρ ($0 < \rho < 1$) positive estimates of the regression parameters are obtained for $\rho \leq 0.94$.

The parameter estimates assume positive values also when the described parametric formulation and the CAR proper model are specified. These
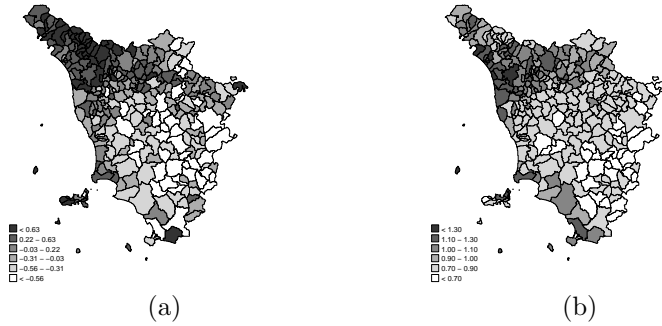
FIGURE 1. Spatial distribution of the education score in 1951 (a) and estimated relative risks 1971-1999 (b).

TABLE 1. β coefficients under different prior assumptions on the spatial random term v_i and their EPD values.

Period	ICAR ($\rho = 1$)	ICAR ($\rho = 0.94$)	CAR proper	Parametric model	Heterogeneity model
1971-74	-0.174	0.187	0.184	0.194	0.194
	(-0.221,-0.134)	(0.134,0.241)	(0.129,0.241)	(0.140,0.2531)	(0.141,0.251)
1975-79	-0.149	0.138	0.133	0.147	0.149
	(-0.189,-0.113)	(0.095,0.184)	(0.089,0.179)	(0.103,0.196)	(0.102,0.198)
1980-84	-0.084	0.106	0.102	0.126	0.127
	(-0.115,-0.052)	(0.067,0.147)	(0.065,0.140)	(0.084,0.172)	(0.084,0.174)
1985-89	-0.067	0.051	0.050	0.073	0.073
	(-0.108,-0.026)	(0.009,0.051)	(0.015,0.093)	(0.039,0.115)	(0.039,0.112)
1990-94	-0.084	0.035	0.035	0.059	0.060
	(-0.131,-0.035)	(-0.003,0.073)	(-0.006,0.069)	(0.026,0.093)	(0.027,0.093)
1995-99	-0.042	0.071	0.072	0.092	0.093
	(-0.096,0.015)	(0.039,0.106)	(0.042,0.103)	(0.060,0.124)	(0.061,0.127)
EPD	2132.135	2115.199	2155.294	2120.796	2117.341

last estimates are also very similar to those obtained when model (1) is modified not including the v_i term (heterogeneity model).

The marginal posterior distributions for the parameters of interest are approximated by Monte Carlo Markov Chain methods.

Bayesian model selection using EPD (Table 1) confirms, in part, what we could expect looking at the β's estimates. In fact, despite of the different prior assumptions, the heterogeneity, the ICAR with $\rho = 0.94$ and the parametric models exhibit not only the same values of the regression coefficients but also very similar deviance statistics suggesting no need for the clustering term.

5   Conclusion and discussion

The effect of the spatial dependence may be investigated through the sensitivity of the regression coefficients (and relative standard errors) to different specifications of the prior distribution of the clustering term. Moreover,

the outcomes of a model without the spatially structured component could be used to see if the form of the spatial structure significantly affects the analysis (Wakefield, 2003).

Our results suggest that when covariates and clustering terms show a strong correlation the standard ICAR assumption could be misleading with regard to the strength of association.

On the other hand, the little differences in the results between the heterogeneity model versus the parametric and the CAR proper assumptions could suggest either (i) that the last two specifications for the clustering term result into too strong limits to the "borrow strength" between areas or (ii) that the covariate adequately explains the spatial structure of risk, and there is no need of the clustering term. The EPD values seem to confirm this last hypothesis.

References

Clayton, D.J., Bernardinelli, L., and Montomoli, C. (1993). Spatial Correlation in Ecological Analysis. *Journal of Epidemiology*, **22**, 1193-1202.

Dreassi, E., Biggeri, A., and Catelan, D. (2003) Space-time models with time dependent covariates for the analysis of the temporal lag between socio-economic factors and lung cancer mortality. **Under revision**.

Journel, A.G. and Huijbregts, C.J. (1978) *Mining Geostatistics.* Academic Press, London.

Knorr-Held, L. (2000) Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, **17-18**, 2555–2568.

Laud, P., and Ibrahim, J. (1995) Predictive Model Selection. *Journal of the Royal Statistical Society*, Ser. B, **57**, 247-262.

Stern, HS, Cressie, NA. Inference for extreme in disease mapping. *Disease mapping and risk assessment for public health*, Lawson A, Biggeri A, Bohning D, Lesaffre E, Viel JF, Bertolini R. (Eds.), Chichester: Wiley, 1999; pp 63–84.

Vigotti, MA., Biggeri, A., Dreassi, E., Protti, MA., and Cislaghi, C. (2001) *Atlas of mortality in Tuscany 1971-94*. Edizioni Plus: Università degli studi di Pisa.

Wakefield, J. (2003) Sensitivity Analysis for Ecological Regression. *Biometrics*, **59**, 9-17.

Bayesian focused clustering for a case-control study on lung cancer in Trieste

Annibale Biggeri[1], Emanuela Dreassi[1], Corrado Lagazio[2] and Marco Marchi[1]

[1] Department of Statistics "G. Parenti", University of Florence, Viale Morgagni 59, I-50134 Florence (Italy) email: {abiggeri,dreassi,marchi}@ds.unifi.it
[2] Department of Statistical Science, University of Udine, Via Treppo 18, I-33100 Udine (Italy) email: lagazio@dss.uniud.it

Abstract: The relationship between four putative sources of environmental pollution (incinerator, shipyard, iron foundry and city center) and lung cancer risk for men in Trieste (Italy), is investigated using a Bayesian framework by a case-control study. In the analysis information on smoking habits and exposure to occupational carcinogens are taken into account to adjust for known risk factor as potential confounders. The models are based on distances between subject place of residence and the different sources of environmental pollution, as a proxi for exposure. Models enable estimation of the risk gradient and directional effects separately for each putative source.
We found that risk of lung cancer is highly related to the city center and incinerator sources. However, as the models appeared to be sensitive to modelling choices, any point analysis should be provided with careful sensibility analysis.

Keywords: Case-control study; Focused Clustering; Hierarchical Bayesian Models; Environmental Pollution.

1 Introduction

In the last years there has been increased interest in modelling disease risk in relation to a point source using a Bayesian framework; see, for example, Wakefield and Morris (2001), Lawson *et al.* (2003) and Congdon (2003).
We use a hierarchical Bayesian model for a case-control study. The models are based on distances between subject (case or control) place of residence and the different sources of environmental pollution, as a proxi for exposure. We present analysis of the spatial pattern of risk of lung cancer for males in Trieste (Italy) with regard to four source, shipyard, iron foundry, incinerator and the city center, while adjusting for known risk factors.
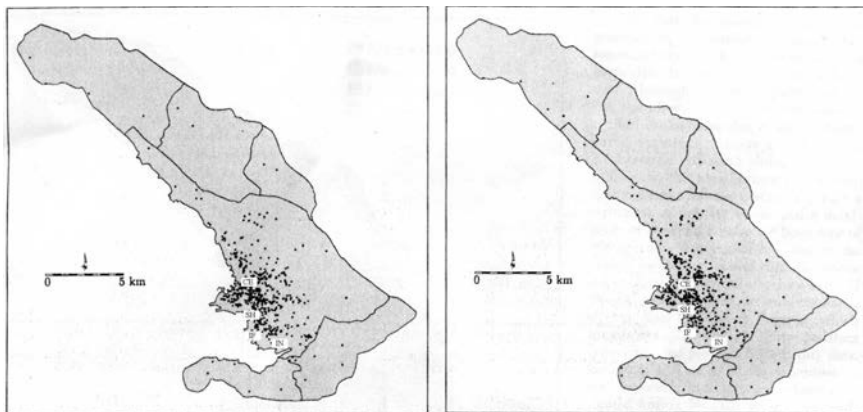
FIGURE 1. Locations of cases (left), controls (right) and putative sources of environmental pollution: city center (ce), shipyard (sh), iron foundry (if) and incinerator (in). Lung cancer males, Trieste (Italy), 1979-1986

2   Data

Data consists in 755 case of lung cancer for males observed from 1979 to 1986 and 755 controls identified through the local autopsy registry (for further details on the study design see Barbone e*t al.*, 1994 and Biggeri *et al.*, 1996). We have considered the distance from subject's last residence to putative source of environmental pollution: city center (ce), incinerator (in), iron foundry (if) and shipyard (sh). Cases, controls and sources of environmental pollution locations are showed in Figure 1.

Covariates, considered in the study as possible confounders, are: smoking habits (nonsmoker, 1-19, 20-39, more than 40 cigarettes per day), exposure to occupational carcinogens (none, possible, likely).

3   The model

A logistic regression model can be defined in terms of odds of having the disease being resident at distance d_s from the source s ($s = ce, in, if, sh$). For subject i ($i = 1, \dots, 1510$) we specify the following logistic model: $Y_i \sim \text{Binomial}(p_i, 1)$,

$$\text{odds}_i = \alpha_0 \prod_j \exp(z_j \gamma_j) \left[1 + \sum_s f(d_{s\,i}) \right] \qquad (1)$$

where α_0 is a constant term, z_j are potential confounders, as smoking habits and exposure to occupational carcinogens, and γ_j the log odds ratio for the

j-th risk factor z_j.

The distance function proposed by Diggle (1990) is the function used in this work

$$f(d_{s\,i}) = \alpha_s \exp(-\beta_s d_{s\,i}) \qquad (2)$$

d_s represents the distance (in meters) from the source of environmental pollution, α_s the excess relative risk at the source location and β_s the exponential decrease of the excess relative risk for longer distance. We have used this distance because it could be extended to include more than one source of environmental pollution simultaneously in the model.

As the distances from the four putative sources are correlated, we chose to consider the city center as part of the model (because the most important source from a statistical point of view) and then assess the significance of the inclusion of each other source in turn.

To allow for directional effects, we define the following distance function for a given source:

$$f(d_{s\,i}; \theta_{s\,i}) = \alpha_s \exp\left[-\beta_s d_{s\,i} + \beta_{s\,sin}\sin(\theta_{s\,i}) + \beta_{s\,cos}\cos(\theta_{s\,i})\right] \qquad (3)$$

where $\theta_{s\,i}$ is the angle between the i-th case or control and source s locations.

Prior distributions Normal(0,10000) are defined for α_0, γ_j, $\beta_{s\,sin}$ and $\beta_{s\,cos}$. Prior for the coefficients relating to the source are $\alpha_s \sim$ Gamma(2, 1) and $\beta_s \sim$ Uniform(0, 1).

We have made use of WinBUGS software (see Spiegelhalter *et al.*, 2000) in order to perform the MCMC analysis. For each model we have run two independent chains; checks for achieved convergence of the algorithm was performed following Gelman and Rubin (1992). We discard the first 100,000 iterations (burn-in) and to store for estimation 5,000 samples.

4   Results and discussion

Coefficient estimates and credibility intervals obtained from the model with only potential confounders are reported in Table 1. Coefficients estimates for the models considering one source of environmental pollution at time are reported in Table 2, for models with two source results are reported in Table 3, for model with directional effect results are reported in Table 4.

Generally speaking results are consistent with previous analysis (Barbone *et al.*, 1994 and Biggeri *et al.*, 1996). All models appeared to be sensitive to modelling choices these suggest that any point analysis should be provided with careful sensibility analysis. Table 5 describes results for different choices of prior distributions for model considering distance from city center and incinerator sources: (a) Congdon (2003)'s priors $\alpha_{ce} \sim$ Gamma(1, 1) and $\alpha_{in} \sim$ Gamma(1, 1), (b) non informative priors $\alpha_{ce} \sim$ Gamma(0.2, 0.1) and $\alpha_{in} \sim$ Gamma(0.2, 0.1), (c) priors based on maximum likelihood estimates $\alpha_{ce} \sim$ Gamma(2, 1) and $\alpha_{in} \sim$ Gamma(7, 1).

TABLE 1. Estimates of coefficients for potential confounders (odds ratio)

Confounder		Estimates(CI 95%)
Smoking	nonsmoker	ref.
	1-19 cigarettes/day	7.393 (4.459,12.580)
	20-39 cigarettes/day	13.571 (8.156,22.750)
	≥ 40 cigarettes/day	22.316 (12.850,38.650)
Occupational exposure	no	ref.
	possible	1.284 (1.003,1.643)
	probable	2.217 (1.634,2.932)

TABLE 2. Estimates of coefficients for the distance from each source of environmental pollution

Source	α_s (CI 95%)	β_s (CI 95%)
City center (ce)	2.560 (0.519,6.194)	0.531 (0.059,0.959)
Shipyard (sh)	1.696 (0.350,4.489)	0.128 (0.008,0.899)
Iron foundry (if)	2.044 (0.412,5.481)	0.282 (0.016,0.926)
Incinerator (in)	2.233 (0.504,5.287)	0.262 (0.009,0.897)

TABLE 3. Estimates of coefficients for the distance from city center and other sources

	sh	if	in
α_{ce}	2.626 (0.423,6.321)	2.561 (0.570,5.893)	2.371 (0.600,5.763)
β_{ce}	0.555 (0.036,0.964)	0.457 (0.015,0.941)	0.369 (0.014,0.908)
α_s	1.402 (0.302,3.642)	2.210 (0.437,5.492)	2.549 (0.579,5.968)
β_s	0.197 (0.009,0.939)	0.263 (0.023,0.918)	0.236 (0.033,0.816)

TABLE 4. Estimates of coefficients for the distance from city center and incinerator sources considering directional effects for incinerator

Coefficient	Estimate (CI 95%)
α_{ce}	2.424 (0.563,5.925)
β_{ce}	0.409 (0.021,0.920)
α_{in}	2.140 (0.414,5.630)
β_{in}	0.292 (0.032,0.877)
$\beta_{in\,sin}$	-0.525 (-2.135,1.041)
$\beta_{in\,cos}$	0.083 (-1.219,1.748)

TABLE 5. Estimates of coefficients for the distance from city center and incinerator sources for several choices of prior distributions. (a) Congdon (2003)'s priors (b) non informative priors (c) priors based on maximum likelihood estimates

Coefficient	Priors		
	(a)	(b)	(c)
	Estimate (CI 95%)	Estimate (CI 95%)	Estimate (CI 95%)
α_{ce}	1.781 (0.223,4.966)	4.275 (0.017,14.96)	2.275 (0.627,5.686)
β_{ce}	0.374 (0.019,0.917)	0.506 (0.019,0.962)	0.306 (0.014,0.892)
α_{in}	1.790 (0.192,4.632)	3.677 (0.001,14.19)	6.645 (2.753,12.18)
β_{in}	0.235 (0.023,0.872)	0.300 (0.028,0.906)	0.335 (0.123,0.856)

References

Barbone, F., Bovenzi, M., Cavallieri, F., and Stanta, G. (1994). Air pollution and Lung Cancer in Trieste, Italy. *American Journal of Epidemiology*, **141**, 1161–1169.

Biggeri, A., Barbone, F., Lagazio, C., Bovenzi, M., Stanta, G. (1996). Air pollution and Lung Cancer in Trieste, Italy: Spatial Analysis of Risk as a Function of Distance from Sources. *Environmental Health Perspectives*, **104**, 7, 750–754.

Congdon, P. (2003). *Applied Bayesian Modelling*, John Wiley & Sons, Ltd.

Diggle, P.J. (1990). A Point Process Modelling Approach to Raised Incidence of a Rare Phenomenon in the Vicinity of a Prespecified Point. *Journal of the Royal Statistical Society A*, **153**, 340–362.

Gelman, A., and Rubin, D.R. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, 457–511.

Lawson, A., Browne, W.J., and Vidal Rodeiro, C.L. (2003). *Disease mapping with WinBugs and MlWin*. Chichester: John Wiley & Sons Ltd.

Spiegelhalter, D.J., Thomas, A., Best, N.G., and Gilks, W.R. (2000). *WinBugs*, Medical Research Council Biostatistics Unit, Cambridge.

Wakefield, J.C., and Morris, S.E. (2001). The Bayesian Modelling of Disease Risk in Relation to a Point Source. *Journal of the American Statistical Association*, **96**, 453, 77–91.

Imputing missing phenotypes:
A new family-based association test

Amy Murphy[1], Deborah Blacker[2], and Christoph Lange[1]

[1] Department of Biostatistics, Harvard School of Public Health, 655 Huntington Ave., Boston, MA 02115, USA, email: amurphy@hsph.harvard.edu

[2] Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, 149 13th Street, Charlestown, MA 02129, and Department of Epidemiology, Harvard School of Public Health, 655 Huntington Ave., Boston,MA 02115

Abstract: We define a new test statistic that accommodates missing phenotypic data in family-based association tests (FBATs). The missing phenotypes are imputed using the conditional mean model (Lange et al. (2003)). When the outcome data are missing at random, FBAT-IMP demonstrates higher power, in both simulations and an Alzheimer study, than the standard quantitative FBAT.

Keywords: Family-Based Association Tests, MCAR, MAR, conditional mean model, Alzheimer disease, time-to-onset

1 Introduction

Family-based association studies of disease outcomes and genetic markers use samples of diseased subjects along with their parents or other family members. Family-based association tests (FBATs) are constructed using the genetic data of the family members to calculate the distribution of the test statistic under the null hypothesis, conditioning on phenotypes and parental genotypes (Rabinowitz and Laird 2000). In studies of diseases with late onset, e.g. Alzheimer disease, the parental genotypes are usually not available and additional siblings must be genotyped to construct the marker distribution. For many late-onset diseases, a typical study design is to ascertain large sib-ships in which at least one sibling is affected. In studying these late onset diseases, genes may be modelled as quantitative trait loci (QTL) for age of onset (Daw et al. 2000). A primary issue in studying gene association with these diseases is how to best utilize the information from offspring that are unaffected. In the standard FBAT statistic, the affected siblings contribute their phenotypic and genetic information to the test statistic, while the unaffected siblings are only used for the computation of the marker distribution under the null-hypothesis. The challenge is constructing an FBAT statistic such that information from offspring unaffected at the time of analysis may also contribute to the statistic.

In this paper, we propose a new test statistic for family-based studies that imputes the missing phenotypic data based on the conditional mean approach (Lange et al. 2003). The imputation is performed under two assumed patterns of missingness, missing completely at random (MCAR) and missing at random (MAR). It can be shown analytically that when the phenotypic data are missing completely at random, no additional power can be obtained by imputing the missing phenotype. However, when the data are missing at random, additional power is achieved by imputation (Murphy et al. 2004a). The magnitude of the power increase is assessed by simulation studies. An application to time-to-onset data from an Alzheimer study shows the practical relevance of our method. Such studies frequently encounter missing at random data, since a genetic variant may delay/accelerate disease onset, thus creating different patterns of missingness.

2   The Data Set: Alzheimer study, Blacker et al.(1998)

The data set is from the NIMH Genetics Initiative Alzheimer Disease sample (Blacker et al. 1998) and has been previously analyzed in Lange et al. (2004). We will re-analyze 2 alleles at the APOE locus. The data set contains 143 nuclear families with 2-10 siblings (Blacker et al. 1998). The parental genotypes are unknown. Within each family, the first sibling always has Alzheimer's disease. Its genotype and time-to-onset are recorded. The additional siblings are either affected or unaffected with either the time-to-onset or the censoring time given. The genotypes of the additional offspring are known.

3   Methods

Although we will analyze data on multiple siblings without parental genotypes, for simplicity, we will derive the methodology using trios, i.e., one offspring per family and the parental genotypes are known. Our methodology extends readily to scenarios in which the parental genotypes are not known, using the approach by Rabinowitz and Laird (2000).
In the study, n independent trios are sampled and a bi-allelic marker locus with alleles A and B is genotyped. We denote the number of transmitted A alleles in the offspring of the ith family by x_i. The parental genotypes in the ith family are p_{i1} and p_{i2}. For each offspring, a quantitative trait, e.g. time-to-onset, is recorded and denoted by y_i. The conditional mean model (Lange et al. 2003) for the ith offspring, is then given by $E(Y_i|p_{i1}, p_{i2}) = a \cdot E(X_i|p_{i1}, p_{i2})$, where a denotes the true additive genetic effect size. For simplicity, we assume here that the offset is 0. The conditional mean model has the advantage that the additive genetic effect size can be estimated using all observed phenotypic data and the parental genotypes without biasing the significance level of any subsequently computed family-based association test (Lange et al. 2003). Next, the conditional mean model is

extended to accommodate missing data. If data is not observed (e.g., onset is censored), we set y_i to missing, i.e., $y_i = \text{NA}$, and denote this observation by $y_{i,mis}$. Denoting the estimate for the genetic effect size by $\hat{a}$, the missing phenotypic data can be imputed by $\hat{y}_{i,mis} = \hat{a} \cdot \text{E}(X_i|p_{i1}, p_{i2})$. We then define the FBAT statistic for observed and imputed data. Using matrix notation, let $\mathbf{Y_{par}} = \left(\hat{\mathbf{Y}}_{\mathbf{mis}}^t, \mathbf{Y}_{\mathbf{obs}}^t\right)^t$. The vector $\mathbf{Y_{par}}$ has been partitioned into observed and missing outcomes, where $\hat{\mathbf{Y}}_{\mathbf{mis}}$ denotes the sub-vector of missing phenotypes that have been imputed using the conditional mean model. The vector of marker alleles, $\mathbf{X}$, and its expected value conditioned upon parental genotypes $\text{E}(\mathbf{X}|\mathbf{P_1 P_2})$ are partitioned in the same manner. Lastly, let $\bar{\mathbf{Y}}_{\mathbf{obs}}$ denote the phenotypic mean among the observed outcomes. Thus, the test statistic FBAT-IMP is given by:

$$S = \mathbf{T}^t \left[\mathbf{X_{par}} - E(\mathbf{X_{par}}|\mathbf{P_1}, \mathbf{P_2})\right] \text{ and } D = \mathbf{T}^t Var(\mathbf{X_{par}})\mathbf{T}, \qquad (1)$$

with $\mathbf{T} = \left(\mathbf{Y_{par}} - \bar{\mathbf{Y}}_{\mathbf{obs}}\right)$. Under the hypothesis of no linkage and no association, FBAT-IMP $= S^2/D \sim \chi_1$ (Murphy et al. 2004a).

The standard quantitative FBAT-statistic (here FBAT-OB) (Laird et al. (2000)) is identical to equation 1 above, except that $\mathbf{X_{par}}$ and $\mathbf{Y_{par}}$ are replaced by $\mathbf{X_{obs}}$ and $\mathbf{Y_{obs}}$, respectively. Only the sub-vectors of $\mathbf{X}$ and $\mathbf{Y}$ corresponding to observed phenotype data are used in calculating the test statistic. Under H_0, FBAT-OB $= S^2/D \sim \chi_1$ (Lange et al. 2003).

The following theorems for the power of FBAT-IMP and FBAT-OB were derived by Murphy et al. (2004a):

Theorem 1: Under the assumptions of Hardy-Weinberg and the missingness of the time-to-onset data completely at random, the power of FBAT-OB and FBAT-IMP are identical.

Theorem 2: Under the assumption that $a > 0$ and $\text{P}(x = 2|Y \text{ is missing}) > \text{P}(x = 1|Y \text{ is missing}) > \text{P}(x = 0|Y \text{ is missing})$, the power of FBAT-IMP is greater than the power of FBAT-OB.

The result of Theorem 2 is of practical importance for time-to-onset data. Candidate genes may delay/accelerate disease onset, creating a monotone missingness pattern for time-to-onset as required by Theorem 2. In such situations, using FBAT-IMP can be advantageous to using only the families with observed time to onset data, i.e., using FBAT-OB.

4.1   Results: Simulation study

We assessed the magnitude of the power difference between FBAT-IMP and FBAT-OB by simulation studies. The genetic data was generated using Binomial distributions and Mendelian transmissions. The phenotypic data was simulation by a Normal distribution, using an additive mode of inheritance, i.e., $Y \sim N(ax, 1)$, where a is the additive effect for phenotype and x is the observed number of alleles at the marker locus. To simulate the 'observed' data set, none of the outcomes were removed if zero alleles were

present at the marker locus, 30% were randomly deleted if one allele was missing,and 60% were deleted at random if two protective were present.

For a variety of scenarios, Table 1 displays the estimated power levels. At every allele frequency and heritability level, FBAT-IMP demonstrated greater power than FBAT-OB. Among the higher allele frequencies (10, 20, and 40%), the power of the FBAT-OB levels off, as the increase in allele frequency is offset by the increased missingness, while the power of FBAT-IMP continues to improve. The relative change in power estimates ranges from 15%-40%, with the greatest differences observed at the lowest heritability levels and highest allele frequencies.

4.2   Results: Data analysis

The results of the analysis of the Alzheimer data set (Blacker et al. 1998) are shown in Table 2. Time-to-onset was assumed to be the quantitative trait of interest. Both FBAT-OB and FBAT-IMP detected an association between the marker alleles and time-of-onset of Alzheimer disease. However, in both alleles, FBAT-IMP provided a more significant result. Additionally, to estimate the power of the FBAT-OB and FBAT-IMP statistics calculated for these data, a simulation using the missingness patterns and allele frequencies observed in the alzheimer data set was performed. The frequency of APOE allele 4 was 43%, and the percent missing for 0,1, and 2 alleles was 19, 26, and 28%, respectively. The APOE allele 3 comprised 53% of the observed alleles, with 30, 25, and 16% missing for 0,1, and 2 alleles, respectively. As shown in Table 2, the estimated power universally increases. The increase is modest in the APOE 3 allele, but the power gains seen in the APOE 4 allele are comparable with the simulation study. Despite a finer missingness gradient, the FBAT-IMP still outperforms FBAT-OB.

5   Discussion

In this paper, we presented a new test for family-based association tests when missing data are present. When data are missing at random, FBAT-IMP demonstrates an increase in power over the quantitative FBAT approach, which is particularly useful in complex diseases where the missingness pattern of the phenotype data may be attributable to genetic effects. The power gains are most pronounced at higher allele frequencies and lower heritability levels. We also showed that the power gains are still considerable even when the missingness percentages are more similar across covariate levels. Further testing of the this methodology will involve its extension to multivariate data (Murphy et al. 2004b) As the number of phenotypes increases, missing data issues are far more frequently encountered (i.e., more difficult to ascertain phenotypes), but the missingness pattern is less likely due to genetic reasons.

TABLE 1. Simulation Study-Estimated Power Levels for 1000 trios, $\alpha = .05$

Allele	heritability=0.01		heritability=0.025	
frequency	FBAT-OB	FBAT-IMP	FBAT-OB	FBAT-IMP
0.01	0.34	0.45	0.61	0.75
0.05	0.38	0.50	0.74	0.87
0.10	0.39	0.55	0.75	0.90
0.20	0.39	0.57	0.76	0.92
0.40	0.39	0.63	0.76	0.95

TABLE 2. Association between time-to-onset and APOE-alleles (h = heritability)

Allele	Test Statistic	FBAT	p-value	$\hat{a}$	Power (h=0.01)
3	FBAT-OB	16.26	5.52e-05		0.46
	FBAT-IMP	**19.90**	**8.18e-06**	**3.04 yrs**	**0.49**
4	FBAT-OB	24.17	8.84e-07		0.36
	FBAT-IMP	**26.84**	**2.21e-07**	**-3.08 yrs**	**0.60**

References

Blacker D., Wilcox M.A., Laird N.M., Rodes L., Horvath S.M., Go R.C., Perry R., Watson B.Jr., Bassett S.S., McInnis M.G., Albert M.S., Hyman B.T., Tanzi R.E.(1998). Alpha-2 macroglobulin is genetically associated with Alzheimer disease. *Nature Genetics*, **19**, 357-360.

Daw E.W., Payami H, Nemens E.J., Nochlin D., Wijsman E.M., Bird T.D., Schellenberg G.D. (2000). The number of trait loci in late-onset Alzheimer's disease. *American Journal of Human Genetics*, **66**, 196-204.

Laird N.M., Horvath S., Xu X. (2000). Implementing a unified approach to family based tests of association. *Genetic Epi.*, **19**, S36-S42.

Lange C., Blacker D., Laird N.M. (2004). Family-based association tests for survival and times-to-onset analysis. *Stat. in Medicine*, **23**, 179-89.

Murphy, A.J., and Lange, C. (2004a). Analytical power calculations and asymptotic properties of FBAT-IMP. *Technical Report*.

Murphy, A.J., van Steen, K., Lange, C. (2004b). On missing phenotype data in multivariate family based association tests: FBAT-GEE-IMP and imputation strategies based on the EM-algorithm, the DA-algorithm and the conditional mean model (submitted).

Rabinowitz D., Laird N.M. (2000). A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Human Heredity*, **50**, 211-223.

Conditional Akaike Information for Mixed Effects Models

Florin Vaida [1] and Suzette Blanchard [2]

[1] Department of Biostatistics, Harvard School of Public Health, Boston USA; vaida@sdac.harvard.edu
[2] Frontier Science and Technology Research Foundation Inc., Chestnut Hill, MA 02467, USA; suzette@sdac.harvard.edu.

Abstract: We show that for a linear mixed effects model where the question of interest concerns cluster-specific inference the commonly-used definition for AIC is not appropriate. We propose a new definition for this context, which we call the conditional Akaike information criterion (cAIC). The cAIC is obtained from first principles, and we show that the penalty for the random effects is related to the effective number of parameters ρ proposed by Hodges and Sargent (2001); ρ reflects a level of complexity between a fixed-effects model with no cluster effects, and a corresponding model with fixed cluster-specific effects. We provide finite-sample results for known random effects variances, and an asymptotic approximation for a special case with unknown random effects variances. We compare the conditional AIC with the marginal AIC (in current standard use), and we argue that the latter is only appropriate when the inference is focused on the marginal, population-level parameters. A pharmacokinetics data application is used to illuminate the distinction between the two inference settings, and the usefulness of the conditional AIC.

1   Introduction

Model assessment and comparison are essential aspects of statistical inference. The AIC is, together with the likelihood ratio test, one of the main instruments for model selection. When the model under consideration contains random effects, the definition of the AIC is not straightforward. What likelihood should be used? Should the random effects be counted as parameters or not? In this paper we argue that the answer to these questions depends on the focus of the research question. We distinguish population, or marginal inference, and cluster-specific, or conditional inference. Accordingly, we show that the AIC will be different in the two cases. The formula is the usual one: $\text{AIC} = -2 \log \text{likelihood} + 2K$, where K is the "degrees of freedom" correction, or the number of parameters in the model. However, for the marginal model, the likelihood is the marginal likelihood, and K

is the number of fixed parameters (fixed mean parameters and variance components), whereas for the conditional model, the likelihood is the conditional likelihood (with the random effects at their estimated values), and K is based on the number of effective mean parameters, ρ. Asymptotically, for known variance of the random effects, $K = \rho + 1$, where 1 stands for the unknown error variance σ^2. Spiegelhalter et al. (2002), make an implicit distinction between conditional and marginal inference using the idea of focus of inference for hierarchical models. For hierarchical models, their DIC criterion, based on Bayesian arguments, is also closely related to our conditional AIC.

2   Conditional and marginal linear mixed models

The AIC (Akaike 1973, deLeeuw 1992) is based on the Kullback-Leibler distance $I(f, g) = E_f \log f(y) - E_f \log g(y)$ between the true density f of the distribution generating the data y, and the approximating model for fitting the data, $g(\cdot|\theta)$, $\theta \in \Theta$. This leads to the Akaike information,

$$AI = -2E_{f(y)}E_{f(y^*)} \log g(y^*|\hat{\theta}(y)), \tag{1}$$

which incorporates the model prediction ability of the model g (y^* and y are independent and with same distribution f). When $\hat{\theta}(y)$ is the maximum likelihood estimator (MLE) and the approximating class of models $\mathcal{G}$ is "close" to f, an asymptotic approximation of AI is the Akaike information criterion, AIC $= -2 \log g(y|\hat{\theta}(y)) + 2K$; $K = df$, the number of free parameters in the model $\mathcal{G}$ (Akaike 1973, Burnham and Andersen 2002). A second-order approximation AIC$_c$ yields $K = N(N - df - 1)^{-1}df$, where N is the total sample size (Hurvitch and Tsai, 1989).

Consider a data vector y consisting of observations from m clusters, modeled by the Laird-Ware model (Laird and Ware, 1982) $y_i = X_i\beta + Z_ib_i + \epsilon_i$, $b_i \overset{iid}{\sim} N(0, G)$ where $i = 1, \ldots, m$ is the cluster index, y_i is the vector of n_i responses for cluster i, β is the p-vector of fixed effects, b_i is the q-vector of random effects for cluster i, X_i and Z_i are the $n_i \times p$ and $n_i \times q$ matrices of covariates for the fixed and random effects respectively, and ϵ_i is the error vector. The total number of observations is $N = \sum_{i=1}^{m} n_i$. The errors are independent and normally distributed $\epsilon_i \sim N(0, \sigma^2 I_{n_i})$, independent of the b_i's. The variance matrix G is $q \times q$ and positive semi-definite. In a more condensed notation we write $y = X\beta + Zb + \epsilon$, $b \overset{iid}{\sim} N(0, G_0)$.

Let θ be the vector of parameters in the model, including β, σ^2, and the parameters in the variance matrix G. Conditional on b_i, the likelihood of the model is $g(y|b, \beta, \sigma^2)$, and the marginal likelihood is $g(y \mid \theta) = \int g(y \mid b, \beta, \sigma^2)p(b \mid G)\, db$, where $p(b|G) = \prod_{i=1}^{m} p(b_i \mid G)$ is the distribution of the random effects.

In the Laird-Ware mixed model inference can be made on two levels: (i) population, or marginal inference, and (ii) cluster-specific, or conditional inference. At the population level, the interest lies exclusively in the fixed effects (e.g. the population-averaged treatment effect in a clinical trial) and the marginal mean $E(y_i) = X_i\beta$, whereas the random effects are viewed simply as a way of modeling within-cluster correlation, and therefore are part of the error term $\gamma_i = Z_i b_i + \epsilon_i$. The appropriate AIC here is the usual one, is the one which we call the marginal AIC: mAIC $= -2\log g(\mathbf{y}|\hat{\boldsymbol{\beta}}, \hat{G}, \hat{\sigma}^2) + 2K$, where the likelihood is the marginal likelihood, K is the number of parameters in β, G and σ^2. In contrast, at the cluster level the cluster-specific parameters b_i are of interest themselves, to a great extent they act as parameters, and they are part of the conditional mean $E(y_i|b_i) = X_i\beta + Z_i b_i$. In this case we recommend the cAIC.

3   Conditional AIC for linear mixed effects models

In analogy with (1), we define the conditional Akaike information as

$$\text{cAI} = -2E_{f(y,u)}E_{f(y^*|u)}\log g(y^*|\hat{\theta}(y), \hat{b}(y)) \qquad (2)$$

where the notation is as in (1). For simplicity, assume that the true distribution of y, $f(\cdot|u)$, and $g(\cdot|\theta, b)$ follow the same Laird-Ware model. In addition, u are the true random effects (the realized values which generated the data y), and b are the random effects in the model; $y^*, y \overset{iid}{\sim} f(\cdot|u)$. Given θ, b, the suitable Kullback-Leibler distance between $f(y|u)$ and the model $g(y|\theta, b)$, properly standardized, is $-2E_{f(y|u)}\log g(y|\theta, b)$. The relevant distribution is the conditional. When θ, b are estimated from the data, this measure becomes $-2E_{f(y^*|u)}\log g(y^*|\hat{\theta}(y), \hat{b}(y))$. The measure is evaluated over all possible observed data (y, u), which gives (2). Note that the distribution is conditional for the inner expectation, joint for the outer.

In analogy with the AIC, cAIC is the estimator of the cAI, and is given by the following two results. We assume that the true distribution $f(\cdot|u)$ of the observed data y is given by the Laird-Ware model.

Theorem 1: σ^2, G known. If the variance parameters σ^2 and G are known, an unbiased estimator of the conditional Akaike information is

$$\text{cAIC} = -2\log g(y|\hat{\beta}(y), \hat{b}(y)) + 2\rho. \qquad (3)$$

Here $\hat{\beta}$ is the MLE, and $\hat{b}$ is the empirical Bayes estimator of b.

Theorem 2: σ^2 unknown. Assume that σ^2 is unknown, but $\sigma^{-2}G$ is known. An unbiased estimator of the conditional Akaike Information is

$$\text{cAIC} = -2\log g(y|\hat{\beta}(y), \hat{b}(y)) + 2K$$

where

$$K = \frac{N(N-p-1)}{(N-p)(N-p-2)}(\rho+1) + \frac{N(p+1)}{(N-p)(N-p-2)} \qquad (4)$$

The properties of K are summarized in the following result:

Proposition:
(i) An alternative formula for K is

$$K = \frac{N}{N-p-2}\left[(\rho+1) - \frac{\rho-p}{N-p}\right]$$

(ii)

$$\frac{N(N-p-1)}{(N-p)(N-p-2)}(\rho+1) \leq K \leq \frac{N}{N-p-2}(\rho+1)$$

(iii) As $N \to \infty$, $K/(\rho+1) \to 1$.

Point (iii) states that for large sample sizes $K \approx (\rho+1)$, i.e., counting the degrees of freedom ρ for the mean term and 1 for σ^2. The difference between K and $\rho+1$ is the small sample bias correction (similar to the difference between AIC_c and AIC. These cAIC measures are unbiased for finite samples, not only asymptotically.

4 Application to a Pharmacokinetics Dataset

We analyzed as a case study a pharmacokinetics dataset, the cadralazine data (Lunn, Wakefield et al, 1999). The dataset consists of plasma drug concentrations from 10 cardiac failure patients who were given a single intravenous dose of 30 mg of cadralazine, an anti-hypertensive drug. Each subject has the plasma drug concentration (mg/L) measured at 2,4,6,8,10, and 24 hours, for a total of 6 observations per subject. The data for a given subject are well described by a pharmacokinetic one-compartment model Concentration $= \frac{\text{dose}}{V_d} \times \exp(-k \cdot t)$, where Concentration is the drug concentration at time t, dose is the original dose of the drug (30 mg), V_d is the volume of distribution, and k is the elimination rate constant; V_d and k are the unknown parameters. This corresponds to the linear model $\log(\text{Concentration}) - \log(\text{dose}) = -\log(V_d) - k \cdot t + \text{error}$, written as $y_{ij} = \beta_{0i} + \beta_{1i} \cdot t_j + \epsilon_{ij}$, where $i = 1, \ldots, 10$ stands for the subject, and $j = 1, \ldots, 6$ is the measurement index for subject i. The data for each patient are well described by a straight line, but the slopes and intercepts of the ten regression lines differ from subject to subject. A main interest of the analysis is in determining the distribution log-volume, $-\beta_{1i}$ and elimination rate constants, $-\beta_{2i}$ of the 10 subjects in the study, and their population-level averages. We compare the following two models: 1. Subject-specific

linear regression, β_{0i}, β_{1i} are different, unconstrained parameters for $i = 1, \ldots, m$. 2. Random intercept and slope, i.e. $\beta_{0i} = \beta_0 + b_{0i}, \beta_{1i} = \beta_1 + b_{1i}, (b_{1i}, b_{2i}) \overset{iid}{\sim} N(0, G)$.

The estimators for the linear regression slopes and intercepts are similar for the two models (not included). Based on the parameter estimates and the residuals plot (not included), both models give a very similar fit. We expected the two models to have comparable AIC values. We obtained and AIC of 12.6 for the random effects model, and of -47.1 for the linear regression model. This large difference is not supported by the similar model fit, and by the presumed parsimony advantage of the mixed effects model. In contrast, the asymptotic conditional AIC using $K = \rho + 1$ is -44.5, making the models comparable. The finite sample correction gives even more interesting results for this small-sample dataset: $\mathrm{AIC}_c = -22.8$, and cAIC using (4) is -42.3.

In the appropriate comparison using cAIC, the random effects model is clearly superior.

References

Akaike (1973). Information theory and an extension of the maximum likelihood principle. In *Breakthroughs in statistics (1992)*, vol. 1, 610–624. Springer-Verlag.

Burnham and Anderson (2002). *Model Selection and Multimodel Inference: A Practical Information - Theoretic Approach*. Springer, 2nd edn.

deLeeuw (1992). Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle. In *Breakthroughs in statistics*, vol. 1, 599–609. Springer-Verlag.

Hodges and Sargent (2001). Counting degrees of freedom in hierarchical and other richly parameterized models. *Biometrika* **88**, 2, 367–279.

Hurvich and Tsai (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

Laird and Ware (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

Lunn, Wakefield, Andrew, Best, and Spiegelhalter (1999). *PKBugs Users Guide*. Imperial College of Science, Technology and Medicine, London.

Spiegelhalter, Best, Carlin, and van der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B* **64**, 1–34.

Measuring noncompliance in insurance benefit regulations with randomized response methods for multiple items

Ulf Böckenholt[1] and Peter G.M. van der Heijden[2]

[1] Faculty of Management, McGill University, 1001 Sherbrooke Street West, Montreal, QC H3A 1G5, CANADA

[2] Department of Methodology and Statistics, Utrecht University, P.O. Box 80.140, 3508 TC Utrecht, The Netherlands

Abstract: Randomized response (RR) is a well known method for measuring sensitive behavior. Yet it is not often applied. Two possible reasons for this are (i) its lower efficiency and the resulting need for larger sample sizes, making applications of RR expensive, (ii) the notion that in many applications the RR design may not be followed by every respondent ('cheating').

This paper addresses the efficiency problem by proposing item response theory (IRT) models for the analysis of multivariate RR data. In these models a person parameter is estimated based on multiple measures of a sensitive behavior under study which yields a more efficient and powerful analysis of individual differences than available from univariate RR data. Cheating in a RR study is approached by introducing additional mixture components in the IRT models with one component consisting of respondents who answer truthfully and other components consisting of respondents who do not provide truthful responses to all or a subset of the items.

The resulting IRT model is applied to data from a Dutch survey conducted under receivers of disablement insurance benefit (DIB) who are interviewed about their compliance behavior to rules that are a prerequisite for receiving DIB.

Keywords: randomized response; item response theory; cheating; sensitive behavior; efficiency.

1 Introduction

In many RR studies, respondents are asked multiple questions about one or more domains. For example, in the 2002 surveys conducted in the Netherlands on social security regulation infringements of the Occupational Disability Insurance Act, the Unemployment Insurance Act and the National Social Security Assistance Act, each social security recipient was asked about nine randomized–response questions about their compliance with these regulations. The following four questions focussed on health–related issues: (1) Have you been told by your physician about a reduction in your

disability symptoms without reporting this improvement to your social welfare agency? (2) On you last spot–check by the social welfare agency, did you pretend to be in poorer health than you actually were? (3) Have you noticed personally any recovery from you disability complaints without reporting it to the social welfare agency? (4) Have you felt for some time now to be substantially stronger and healthier and able to work more hours, without reporting any improvement to the social welfare agency?

Clearly, these questions are ordered according to their degree of intentional violations of the regulations. A person who does not report the outcome of a medical investigation may also avoid reporting any personally noticed improvements of their health status. In contrast, persons who notice personal improvements may or may not mis-report their health status. Item–response models (van der Linden et al., 1997) are well–suited for studying how individuals differ in their compliance behavior by ordering respondents on a latent continuum that represents their level of compliance.

Although there is much empirical support to indicate that RR methods increase the number of honest responses, there is no guarantee that all respondents provide truthful answers (see van der Heijden et al, 2000). Some respondents might violate the rules set out by the RR procedure. Here the Forced Choice response format is used: respondents are asked to throw two dice, to answer "yes" when the outcome is 2, 3 or 4, to answer "no" when the outcome is 11 or 12, and to answer truthfully when the outcome is between 5 and 10. A typical rule violation is to answer "no" whatever the outcome of the dice (compare van den Hout et al., 2004; Clark et al., 1998).

To accommodate such response behavior, an extension of the item response approach is presented which allows explicitly for a response bias in the sense that it can capture a possible tendency of respondents towards giving a "No" response regardless of the outcome of the randomizing device. These respondents are captured by a latent class that can be identified by an extreme use of "No" responses.

2   RR Models for Multiple Items

We distinguish three classes of RR models for multiple items. The first class assumes that respondents are homogenous in their compliance behavior and have a fixed probability of answering each item. This is the classical RR model and it is used as a benchmark for the models proposed next. The second model class relaxes the homogeneity assumption and allows for individual variability in compliance for the various behaviors under study. The third class of models considers the possibility that a subset of respondents may not follow the randomization instructions and answer "No" regardless of the outcome of the randomization device.

When all respondents have the same probability of endorsing an item, it

is convenient to express the probability of answering affirmatively by the logistic function with

$$\Pr(x_{ij} = 1) = \Pr(\gamma_j) = \frac{1}{1 + \exp(\gamma_j)}$$

Under random sampling of the respondents, the likelihood function of the homogeneous–response model can then be written as

$$L = \prod_{i=1}^{n} \prod_{j=1}^{J} [\frac{1}{6} + .75 \Pr(\gamma_j)]^{x_{ij}} \ [1 - (\frac{1}{6} + .75 \Pr(\gamma_j))]^{1-x_{ij}}. \tag{1}$$

where $\frac{1}{6}$ is the probability of a forced "yes" and .75 is the probability of a truthful answer. Clearly, the assumption that all respondents have an equal probability of answering an item is too strong in most applications although it is the standard assumption for single-item RR studies.

The second class of models assumes that associations among the responses to multiple items are caused by a person–specific compliance parameter. Because typically the number of items is small in a RR study, we adopt the Rasch (1980) model to measure individual differences in compliance behavior. Under this model, the probability that item j is answered affirmatively by person i can be written as

$$\Pr(x_{ij} = 1) = \Pr(\gamma_j, \theta_i) = \frac{1}{1 + \exp(\theta_i - \gamma_j)},$$

where γ_j is called the item location parameter. Typically, the person parameter θ_i is specified to vary according to a normal distribution.

Under the Forced Choice response format, the item-response model needs to be modified to account for the randomization effect. In this case the likelihood function can be written as:

$$
\begin{aligned}
L \ &= \ \prod_{i=1}^{n} \int \prod_{j=1}^{J} [\frac{1}{6} + .75 \Pr(\gamma_j, \theta_i)]^{x_{ij}} \times \\
&\qquad [1 - (\frac{1}{6} + .75 \Pr(\gamma_j, \theta_i))]^{1-x_{ij}} f(\theta; \mu, \sigma) d\theta, \tag{2}
\end{aligned}
$$

where $f(\theta; \mu, \sigma)$ is the normal density with parameters μ and σ. Note that the mean μ of the population distribution cannot be estimated independently of the item locations. In the reported application, we therefore set $\mu = 0$. It is worthwhile stressing that the normal distribution assumption may not always be appropriate in RR studies and that other distributional forms should be considered to capture more closely the non–compliance variability in the population of interest.

The third class of models allows for the possibility that not all respondents comply with the randomization response format and provide a "No" response regardless of the question asked. Combined with the item–response

model given by (2), the likelihood function is specified as:

$$
L \;=\; \prod_{i=1}^{n} (\pi \int \prod_{j=1}^{J} \{ [\tfrac{1}{6} + .75 \Pr(\gamma_j, \theta_i)]^{x_{ij}} \; [1 - (\tfrac{1}{6} + .75 \Pr(\gamma_j, \theta_i))]^{1-x_{ij}} \}
$$

$$
\times \quad f(\theta; \mu, \sigma) d\theta + (1-\pi) \prod_{j=1}^{J} \{ \Pr(\text{``No''})^{x_{ij}} \; [1 - \Pr(\text{``No''})]^{1-x_{ij}} \}), \quad (3)
$$

where π denotes the probability of a randomly sampled person to answer the questions according to the FC mechanism. In the reported application, we specify that participants who answer "No" regardless of the question asked, give this response with probability 1. It is straightforward to relax this assumption and to estimate the probability of a "No"– response from the data. The crucial assumption of (3) is that members of the "No"-group do not provide any information about the items' location and discrimination parameters.

3   Data Analysis

The aim of the study and the RR design have been described above. We note that 44% of all respondents respondents provide "No" responses to all four items.

The homogeneous model required the estimation of four item location parameters and yielded a goodness-of-fit statistic of $G^2 = 123.8$ with 11 d.f.. Clearly, the assumption of no individual differences does not agree with the data. This result is supported by the fit improvement obtained from Model (2). With one additional parameter, the variance of the normal distribution σ^2, Model (2) provides a major fit improvement ($G^2 = 23.4$ with 10 d.f.). However, despite the better fit, this model does not describe the data satisfactorily. The main reason for the misfit is that the outcome of consistent "No"–responses to the four items is greatly underestimated by (2). Model (3) can address this problem by allowing for the possibility that some respondents select the "No"–response for reasons that are unrelated to the compliance parameter θ. The resulting fit improvement provides support for this specification ($G^2 = 14.3$ with 9 d.f.).

Table 1 contains the corresponding parameter estimates of the three models. We note that the standard errors of Model (1) are too small since this model does not reflect the dependencies among the four responses. In contrast, Model (2) overestimates strongly the degree of heterogeneity in the data since it tries to fit the large percentage of "No"–responses to the four items. Model (3) yields a much reduced but still substantial estimate of the population standard deviation ($\hat{\sigma} = 2.07$). About 196 or 12% of the respondents are classified as consistent "No"–sayers. For the remaining 88% of the respondents, the items are ordered but far away from the mean of

the population distribution. Clearly, a positive response to any of the four items is low at the mean of the population distribution.

TABLE 1. Parameter Estimates (and Standard Errors) of RR–Models for Multiple Items

Parameter	Model (1)	Model (2)	Model (3)
$\hat{\gamma}_1$	3.77 (.56)	9.10 (2.74)	4.56 (.88)
$\hat{\gamma}_2$	3.07 (.30)	8.44 (2.73)	3.99 (.80)
$\hat{\gamma}_3$	2.58 (.20)	7.61 (2.72)	3.42 (.67)
$\hat{\gamma}_4$	1.94 (.13)	5.83 (2.01)	2.63 (.53)
$\hat{\sigma}$	–	4.72 (1.61)	2.15 (.47)
$\ln(\frac{\hat{\pi}}{1-\hat{\pi}})$	–	–	2.07 (.34)

By taking into account that about 12% of the respondents give a "No"-response without providing information about their actual compliance behavior, Model (3) renders more accurate estimates about the compliance rate in the population. Under Model (1) the percentage of non-compliant respondents for the four items are 2.2%, 4.5%, 7.0%, and 12.5% respectively. In contrast, under Model (3) the corresponding estimates are 5.2%, 7.7%, 11.0%, and 17.0%. These differences are substantial and demonstrate the value of the proposed models for the analysis of RR data.

References

Clark, S.J. , and Desharnais, R.A. (1998). Honest answers to embarrassing questions: detecting cheating in the randomized response model. *Psychological Methods*, **3**, 160-168.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: The University of Chicago Press. (Original published 1960, Copenhagen: The Danish Institute of Educational Research)

van der Linden, W. J. and Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory.* New York: Springer.

van den Hout, A. and van der Heijden, P.G.M. (2004). The analysis of multivariate misclassified data with special attention to randomized response data. *Sociological Methods and Research*, **32**, 310-336.

van der Heijden, P.G.M., van Gils, , G., Bouts, J. and Hox, J. (2000). A comparison of randomized response, CASAQ, and direct questioning; eliciting sensitive information in the context of welfare and unemployment benefit. *Sociological Methods and Research*, **28**, 505-537.

Confidence intervals for the variance of random-effects linear models: a new Stata command

Matteo Bottai[1][2] and Nicola Orsini[2][3]

[1] Arnold School of Public Health, University of South Carolina, Columbia, SC 29208, mbottai@gwm.sc.edu
[2] Institute of Information Science and Technology, C.N.R., Pisa, Italy
[3] Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden, nicola.orsini@imm.ki.se

Keywords: boundary; confidence intervals; random effects; score test; variance components.

1 Introduction

The methodological developments presented are implemented in a new Stata command named `xtci`, and an expanded version of the present article is published on the Stata Journal (Bottai and Orsini, 2004).

The random-effects linear model has been widely applied to different areas of data analysis (among many others, Breslow and Calyton 1993, Diggle et al 1994, McCulloch and Searle 2001). In its simplest form, it can be written as

$$y_{it} = \mathbf{x}_{it}^T \beta + u_i + e_{it}, \quad u_i \sim N(0, \sigma_u^2), \quad e_{it} \sim N(0, \sigma_e^2) \tag{1}$$

where y_{it} is the tth observation taken on some random variable Y for the ith unit, with $i = 1, \ldots, m$, $t = 1, \ldots, T_i$; $\mathbf{x}_{it}$ is a covariate vector and β is a parameter vector of fixed effects; u_i is a unit-specific normal random effect with zero mean and variance σ_u^2 that is assumed to be non-negative, and e_{it} is the normal residual error with variance σ_e^2 that is assumed to be strictly positive. Also, u_i and e_{it} are assumed to be independent. Units can refer to individuals on whom repeated observations are taken, or to families whose members are sampled, or to otherwise-defined groups within which observations may be correlated.

In such models it is often of interest to make inference not only about the fixed and random effects but also about the variance components. In particular, testing homogeneity across units is equivalent to testing the null hypothesis

$$H_0 : \sigma_u^2 = 0. \tag{2}$$

In general, testing whether a variance parameter is zero implies testing a parameter value on the boundary of the parameter space, the variance being non-negative. Several authors suggest the use of the large-sample likelihood ratio test that adjusts for the boundary condition. In fact, under this non regular scenario, the asymptotic distribution of the usual likelihood ratio test statistic follows a distribution that is a 50:50 mixture of a $\chi^2_{(1)}$ and the constant zero (Self and Liang 1987). Several statistical packages provide the upper-tail probability of the appropriate asymptotic distribution of the likelihood ratio test statistic.

However, such method cannot be used to construct confidence intervals for the variance of the random effect, σ^2_u. Besides, confidence intervals for the random-effect variance that are based on a Wald-type test, too often used, can be shown to be asymptotically wrong. To the best of our knowledge, no published work has provided methods for constructing likelihood-based confidence regions for the variance component that are asymptotically correct.

It can be shown that inference about the variance component σ^2_u can be accommodated within the non-regular problems of singular information. Such connection had been noted several years ago (Chesher 1984, Lee and Chesher 1986) but only recently a general theory was developed for the singular information case (Rotnitzky et al 2000). Using the results derived for the singular information problem (Bottai 2003), a method is developed and implemented in the Stata command `xtci` that is based on the inversion of a score-type test, which provides asymptotically-correct confidence intervals. Also, when testing the hypothesis of homogeneity across units (2), the proposed method is shown to have better small-sample properties than the one based on the likelihood ratio test statistic.

The remaining sections are organized as follows: Section 2 shows the observed rejection proportions of the confidence intervals generated by `xtci` on simulated data, section 3 presents a real data example and section 4 presents some final remarks.

2   Simulated data

The command `xtci` was applied to simulated data. Three-thousand samples were pseudo-randomly generated for model (1) under a grid of values for the random-effect standard deviation $\sigma_u = 0, 0.01, \ldots, 0.09, 0.10, 10$, and for different numbers of units or groups $m = 10, 100, 1000$. The residual error standard deviation σ_e was set constant to the value one for all the simulation. Two covariates were pseudo-randomly generated from a Uniform$(-1, 1)$ and a Uniform$(0, 2)$ distribution respectively, with $\beta = (1, 2)^T$. The observed rejection proportions over the simulated samples of the 95% confidence intervals provided by the command `xtci` are shown in table 1. For the samples generated under the value $\sigma_u = 0$, the observed

rejection proportions of the likelihood ratio test adjusted for boundary condition at the 0.05-level is also reported.

TABLE 1. Observed rejection proportions of the proposed score-type test and of the likelihood ratio test adjusted for boundary condition among 3000 simulated samples generated under different values of σ_u and number of units or groups for the random-effects linear model (1). (Simulation error $\pm 0.78\%$.)

σ_u	$m=10$	$m=100$	$m=1000$
xtci			
0.00	5.20	5.23	4.63
0.01	5.17	5.43	5.37
0.02	5.03	5.23	4.93
0.03	5.33	5.60	4.57
0.04	5.30	5.07	5.63
0.05	4.73	5.63	5.00
0.06	5.77	5.17	4.93
0.07	5.30	5.63	5.30
0.08	5.27	5.40	4.53
0.09	5.47	5.43	5.30
0.10	4.80	5.20	4.07
10.0	4.57	5.03	4.90
xtreg			
0.00	2.43	4.13	4.27

Regardless of the number of units or groups, m, the observed rejection proportion is close to its nominal level of 5% uniformly across the values of the standard deviation σ_u. Although based on a large-sample test, the command xtci shows acceptable behavior in small samples as well.

The adjusted likelihood ratio test provided by the command xtreg was applied only to the sampled simulated under the value $\sigma_u = 0$. In the present simulation, when the number of units or groups $m = 10$, its observed rejection proportion is 2.43%, well below its nominal level of 5%. In other extensive simulation experiments not reported here, we observed that the rejection proportion becomes satisfactorily close to the nominal level only when the number of units or groups is no smaller than a thousand.

The observed rejection proportion of the confidence regions obtained by inverting the Wald-type test, as provided by the command xtreg, is wrong in small as well as large samples. Depending on the values of σ_u and m, its rejection probability can be as high as 15% or as low as 0.5%. Besides, its confidence intervals may happen to include negative values, which are out

TABLE 2. Maximum likelihood estimates and 95% confidence intervals for the linear random-effects model.

Parameter	Estimate	95% Conf. Int.	
Intercept	2.132	1.654	2.609
Sex (F vs. M)	-0.736	-0.978	-0.493
15–29 yrs	0.924	0.386	1.462
30–44 yrs	1.225	0.706	1.744
45–59 yrs	0.830	0.323	1.336
60–74 yrs	0.596	0.003	1.189
75+ yrs	-1.142	-2.447	0.163
σ_e	1.167	1.034	1.300
σ_u	0.432	0.216	0.681

of the feasible space of the variance parameter.

3   Example: Individual daily moving behavior

A survey on daily moving behaviors of the people residing on the territory of the Municipality of Pisa was carried out in October 2002. Data about the trips made in the preceding 24 hours were recorded on 401 individuals from 272 families. The present analysis is aimed at modelling the logarithm of the total distance covered by each individual in one day as a function of sex and age grouped in classes (0-14, 15-29, 30-44, 45-59, 60-74, 75+ years). To account for the potential dependence of the observations within families, random effects are introduced into a linear regression model as follows,

$$\texttt{logdistance}_{it} = \beta_0 + u_i + \beta_1 \texttt{sex}_{it} + \sum_{k=2}^{6} \beta_k \texttt{ageclass}k_{it} + e_{it}$$

with the notation described for model (1), where the variable `logdistance` is the logarithm of the total distance covered, the variable `sex` is 1 for female and 0 for male, `ageclass2` to `ageclass6` are indicator variables, one for each age class with the youngest class omitted. Maximum likelihood estimates are shown in table 2 where the confidence interval for σ_u is estimated by the proposed procedure and the remaining ones are obtained by inverting Wald tests.

Testing homogeneity across families is equivalent to testing the hypothesis (2). The proposed score-type test, which provides asymptotically correct p-values and confidence intervals, suggests to reject the null hypothesis, with a p-value approximately equal to 0.008. Instead, the likelihood ratio test

p-value divided by two, as we are testing parameters on the boundary (Self and Liang 1987), is 0.082, which is above the usual 0.05 rejection cut-off value. As expected, the proposed procedure has greater power. Although routinely applied, Wald-type tests are asymptotically wrong when testing variance parameters that are close to zero.

4   Final remarks

The command `xtci` was implemented from the results presented by Bottai (2003), and it is the only solution for those seeking to construct confidence intervals for the variance component of a random-effects linear regression model. The procedure described by Bottai (2003) can be extended to non-Gaussian random effects model as well as to many other classes of models, such as generalized linear mixed models and frailty models, whose estimation is based on the likelihood function.

References

Bottai, M. (2003). Confidence regions when the Fisher information is zero. *Biometrika*, 90, 1, 73–84.

Bottai, M. and Orsini, N. (2004). Confidence intervals for the variance component of random-effects linear models. *Stata Journal*, in press.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of American Statistical Association*, 88, 9–25.

Chesher, A. (1984). Testing for neglected heterogeneity. *Econometrica*, 52, 4, 865–872.

Diggle, P.J., Liang, K.-Y., and Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Oxford University Press, London.

Lee, L.F. and Chesher, A. (1986). Specification testing when score test statistics are identically zero. *Journal of Econometrics*, 31, 121-149.

McCulloch, C.E. and Searle, S.R. (2001). *Generalized, linear, and mixed models*. Wiley, New York.

Rotnitzky, A., Cox, D.R., Bottai, M., and Robins, J. M. (2000). Likelihood-based asymptotic inference with singular information. *Bernoulli*, 6, 2, 243–284.

Self, S. G. and Liang, K.-Y. (1987). Properties of maximum likelihood estimators and likelihood ratio test under nonstandard conditions. *Journal of American Statistical Association*, 82, 605-610.

Geoadditive Survival Models

Andrea Hennerfeind, Andreas Brezger and Ludwig Fahrmeir[1]

[1] Ludwig-Maximilians-Universität, Department of Statistics, Ludwigstr. 33, D-80539 München, Germany

Abstract: Survival data often contain spatial information, such as the residence. In many cases the impact of such spatial effects on hazard rates is of considerable interest. We propose flexible continuous–time geoadditive models, extending the classical Cox model by augmenting the common linear predictor with a spatial component and nonparametric terms for nonlinear effects of time and continuous covariates. Markov random fields and penalized splines are used as basic building blocks. Inference is fully Bayesian. We apply our approach to data from a case study that aims to estimate the effect of area of residence and further covariates on waiting times to coronary artery bypass graft (CABG).

Keywords: Bayesian hazard rate model; penalized splines; spatial survival data.

1 Introduction

Nonparametric Bayesian survival models have become quite popular in recent years, and some previous work deals with related, special cases of our approach. Ibrahim et al. (2001) provide a very good overview. In this paper modelling and inference is developed from a Bayesian perspective, using information from the full likelihood rather than from a partial likelihood, in combination with priors for parameters and functions. Estimation of unknown functions of time and continuous covariates is based on Bayesian penalized spline (P–spline) regression (Lang and Brezger, 2004). Basically, time is treated in the same way as a continuous covariate, but the degree and amount of smoothness may be different. The spatial component is modelled by Gaussian Markov random field priors.

2 Models, likelihood, and priors

Consider survival data in usual form, i.e., it is assumed that each individual i in the study has a lifetime T_i and a censoring time C_i that are independent random variables. The observed lifetime is then $t_i=\min(T_i, C_i)$, and δ_i denotes the censoring indicator. The data are given by

$$(t_i, \delta_i; v_i), \quad i = 1, \ldots, n$$

where v_i is the vector of covariates.

In Cox's proportional model the hazard rate for individual i is assumed as

$$\lambda_i(t) = \lambda_0(t) \exp(\gamma_1 v_{i1} + \ldots + \gamma_r v_{ir}) = \lambda_0(t) \exp(v_i' \gamma). \tag{1}$$

The baseline hazard rate is unspecified, and, through the exponential link function, the covariates $v = (v_1, \ldots, v_r)$ act multiplicatively on the hazard rate. In a number of applications there is a need for extending this basic model with respect to several aspects. We propose novel nonparametric Bayesian survival models that can deal with these issues in a flexible and unified framework. Reparametrizing the baseline hazard rate through $\exp\{f_0(t)\}$, $f_0(t) = \log\{\lambda_0(t)\}$, partitioning the vector of covariates into groups x, z, and v and adding a spatial index s, we extend model (1) to

$$\lambda_i(t) = \exp(f_0(t) + \sum_{j=1}^{p} f_j(t) z_{ij} + \sum_{j=p+1}^{p+q} f_j(x_{i,j-p}) + f_{spat}(s_i) + v_i' \gamma). \tag{2}$$

Here $f_j(t)$ are time–varying effects of covariates z_j, $f_j(x)$ is the nonlinear effect of a continuous covariate x, $f_{spat}(s)$ is the structured effect of the spatial index s, with $s_i = s$ if unit i is from area s, $s = 1, \ldots, S$, and γ is the vector of usual linear fixed effects.

Under the usual assumption about noninformative censoring, the likelihood is given by

$$L = \prod_{i=1}^{n} \lambda_i(t)^{\delta_i} \cdot \exp\left(-\int_0^{t_i} \lambda_i(u) du\right) = \prod_{i=1}^{n} \lambda_i(t)^{\delta_i} \cdot S_i(t). \tag{3}$$

The Bayesian model formulation is completed by assumptions about priors for parameters and functions. We assume diffuse priors for fixed effect parameters γ. For unknown functions f_j, we assume Bayesian P–spline priors (Lang and Brezger 2004). The idea of P–spline regression is to approximate a function as a linear combination of B–spline basis functions B_m, i.e.

$$f_j(x) = \sum_{m=1}^{M_j} \beta_{jm} B_m(x).$$

The basis functions B_m are B–splines of degree l defined over a grid of equally spaced knots. The number of knots is rather high, to maintain flexibility, but smoothness of the function is encouraged by difference penalties for neighboring coefficients in the sequence $\beta_{j1}, \ldots, \beta_{jM_j}$. The Bayesian analogue are random walk smoothness priors. The amount of penalization is controlled by the variance τ_j^2, which acts as a smoothness parameter.

Considering small area data with sparse data for at least some of the areas, fixed area–specific effects would not lead to reliable estimations. Therefore we fit a structured spatial effect by assuming Markov random field priors.

This technique borrows strength from neighboring areas, i.e. we assume that neighboring areas (i.e. areas that share a common boundary) are more similar than arbitrary areas and therefore the spatial effect varies smoothly. We assume that the effect of an area s is normally distributed

$$f_{spat}(s) := \beta_s^{spat} \sim N(\frac{1}{N_s} \sum_{j \in \delta_s} \beta_j^{spat}, \frac{\tau_s^2}{N_s}),$$

where N_s is the number of neighbors of area s, and $j \in \delta_s$ denotes that area j is a neighbor of area s. The amount of smoothness is controlled by a smoothing parameter τ_s^2 that is estimated jointly with the parameters β_s. Variances τ_j^2 as well as τ_s^2 follow weakly informative inverse Gamma priors. The Bayesian model specification is completed by assuming that all priors for parameters are conditionally independent, and that all priors are mutually independent.

3   Markov Chain Monte Carlo inference

Full Bayesian inference is based on the entire posterior distribution of all parameters given the data, which is proportional to the product of the likelihood and the prior distributions of all parameters.
The likelihood is given by inserting (2) into (3), but integration over all terms depending on survival time t is required, i.e. terms of the form

$$I_i = \int_0^{t_i} \exp(f_0(u) + \sum_{j=1}^{p} f_j(u) z_{ij}) du.$$

Apart from B–splines of degree zero, i.e. random walk models, and linear B–splines, these integrals are not available in closed form. The first case leads to the piecewise exponential model, where the likelihood is proportional to a Poisson–likelihood with an offset term. For linear B–splines, the integrals can still be solved, but the computational effort is quite high. Therefore we use numerical integration in form of the trapezoidal rule for linear B–splines as well as for the commonly used cubic B–splines.
Full Bayesian inference via MCMC simulation is based on updating full conditionals of single parameters or blocks of parameters.
For updating the parameter vectors corresponding to the time–independent functions $f_j(x)$, as well as spatial effects β_s and fixed effects γ, we use a modified version of an MH–algorithm based on iteratively weighted least squares (IWLS) proposals, see Hennerfeind et al. (2003).
For the parameters corresponding to the functions depending on time t, the IWLS–MH algorithm requires considerably more computational effort. Therefore, we adopt a computationally faster MH–algorithm based on conditional prior proposals, although IWLS–MH has better mixing properties. The full conditionals for the variance parameters are inverse gamma and updating can be done by simple Gibbs steps.

a)   b)
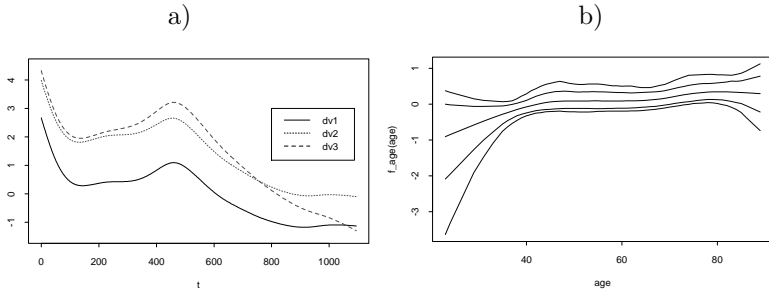
FIGURE 1. a) (log–)baseline effects on time to CABG: posterior mean estimates for 1 diseased vessel (dv1), 2 diseased vessels (dv2) and 3 diseased vessels (dv3) b)Posterior mean estimates of the effect of age and 80% and 95% credible intervals

4   Application: Waiting times to CABG

We illustrate our methods by an application to data from a study in London and Essex that aims to analyze the effects of area of residence and further covariates on waiting times to coronary artery bypass graft (CABG). The data comprise observations for 3015 patients with definite coronary artery disease. Covariates are, among others, sex, age (in years), numbers of diseased vessels (1, 2, 3), and residence (one of 488 electoral wards). The data were previously analyzed by Crook et al. (2003) who classified waiting times in months and applied discrete–time survival methodology. They analyzed and compared a hierarchy of models. Here we apply continuous–time geoadditive survival models, with waiting times given in days as in the original data set, and predictors based on model 12 in Crook et al. (2003), which corresponds to a nonproportional continuous–time model with hazard rate

$$\lambda(t) = \exp(f_0(t) + f_{\text{age}}(\text{age}) + f_s(\text{ward}) + \gamma_1 \text{sex} + f_1(t)\text{dv2} + f_2(t)\text{dv3}), \quad (4)$$

where $f_0(t)$ is the log–baseline rate, $f_{\text{age}}(\text{age})$ is the nonlinear effect of age and $f_s(\text{ward})$ is the structured spatial effect modelled through a MRF prior. The remaining covariates are dummy–coded: sex=1 for female, and sex=0 for male, dv2=1 if the number of diseased vessels=2, dv2=0 else, and dv3=1 if the number of diseased vessels=3, dv3=0 else.
For the (log–) baseline as well as for $f_1(t)$, $f_2(t)$ (the time–varying effects of dv_2 and dv_3) and f_{age} we assumed a cubic P–spline prior with 20 knots. Model (4) can be interpreted as a model with three separate baseline effects $f_0(t)$, $f_0(t) + f_1(t)$, $f_0(t) + f_2(t)$ for patients with one, two or three diseased vessels, respectively. The corresponding estimated curves are displayed in Figure 1a. All baseline effects show an initially high, but strongly decreasing chance of CABG immediately after diagnosis, followed by a slow increase between 150-450 days. Later, the chance of being operated decreases, but
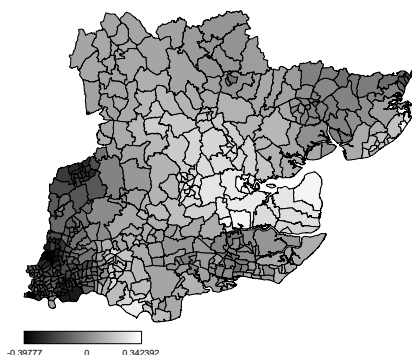
FIGURE 2. Posterior mean estimates of the structured spatial effect

the baseline effect of patients with three diseased vessels decreases more rapidly and crosses the two other curves, which indicates that the proportional hazards assumption is violated. The effect of age (Figure 1b) is almost constant between 40 and 80 years and does not have significant influence. The effect of sex is nonsignificant as well. The map in Figure 2 gives an impression of the spatially varying chance of CABG with light (dark) areas indicating an increased (decreased) effect. Areas with increased chances are Chelmsford and Malden in North Essex, while in some areas in North Essex and North East London patients have to wait longer for surgery.

5   Conclusions

Spatial extensions for analyzing survival data will be of increasing relevance because spatial small–area information is often available. We have developed a flexible class of nonparametric geoadditive survival models within a unified Bayesian framework. Extensions as to more general event history models and censoring mechanisms could be considered in future research.

References

Crook, A., Knorr-Held, L., and Hemingway, H. (2003). Measuring spatial effects in time to event data. *Statistics in Medicine*, **22**, 2943–2961.

Hennerfeind, A., Brezger, A., and Fahrmeir, L. (2003). Geoadditive Survival Models. Discussion Paper 333, SFB 386, University of Munich.

Ibrahim, J.G., Chen, M.H., and Sinha, D. (2001). *Bayesian Survial Analysis*. New York: Springer–Verlag.

Lang, S., and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, **13**, 183–212.

Statistical Models for Market Segmentation

L. Camilleri[1], M. Green[1]

[1] Centre for Applied Statistics, Lancaster University, Lancaster LA1 4YF, UK.

Abstract: It is an essential element of market research that customer preferences are considered and the heterogeneity of these preferences is recognized. By segmenting the market into homogeneous clusters the preferences of customers is addressed. Latent class methodology for conjoint analysis, proposed by Green (2000), is one of the several conjoint segmentation procedures that overcome the limitations of aggregate analysis and prior segmentation. This approach proposes the proportional odds model as a proper statistical model for ordinal categorical data in which the item attributes are included in the linear predictor. The likelihood is maximized through the EM algorithm. This paper considers two extensions of this methodology that incorporate individual characteristics into the models.

Keywords: Proportional Odds Model; Latent Class Model; EM algorithm; Conjoint Analysis; Segmentation.

1 A General Model

Individuals are presented with several items representing different products and are asked to rate each item on an ordinal scale. The observation y_{nj} is a rating response to the jth item elicited by the nth respondent. Consider the Proportional Odds Model as in Green (2000)

$$P\left(y_{jn} = r | \alpha, \beta\right) = F\left(\alpha_r + \mathbf{x}_j' \beta\right) - F\left(\alpha_{r-1} + \mathbf{x}_j' \beta\right)$$

In the first approach this is extended to include individual characteristics together with item attributes in the same linear predictor.

$$P\left(y_{nj} = r | \alpha, \beta\right) = F\left(\alpha_r + \eta_{nj}\right) - F\left(\alpha_{r-1} + \eta_{nj}\right)$$

β is a vector of regression parameters and α is a vector of cut-point parameters. The linear predictor $\eta_{nj} = \eta\left(\mathbf{x}_j, \mathbf{z}_n\right)$ includes item attribute covariates, $\mathbf{x}_j$, individual covariates, $\mathbf{z}_n$ and interaction terms. In market research η_{nj} is referred to as the worth or utility. The choice of $F(.)$ considered is the extreme value distribution leading to the complementary log-log link. The proportional odds model assumes that all respondents act in a similar way in their choice behaviour and that it treats all respondents as homogeneous. One of the criteria for effective market segmentation is to identify differences between distinct groups of customers in the market and

the ability to classify each customer into a segment. For the segmentation procedure a latent class model with K segments is considered.

$$P\left(\mathbf{y}_{nj} = \mathbf{r}|\alpha, \beta, \pi\right) = \sum_{k=1}^{K} \pi_k . P\left(\mathbf{y}_{nj} = \mathbf{r}|\alpha, \beta_k\right)$$

where π_k is the proportion of respondents in the kth segment and the parameters within the segments are estimated at the same time that the segments are uncovered.

In the second approach only the item attributes are included in the proportional odds model as in (Green,2000). The individual characteristics are now included in a mixture model through a classifying function π_{nk}. The choice of parameterization for π_{nk} corresponds to a multinomial logit probability model.

$$\pi_{nk} = \frac{\exp\left(\mathbf{z}_n' \gamma_k\right)}{\sum_{k=1}^{K} \exp\left(\mathbf{z}_n' \gamma_k\right)}$$

The mixed model blends this multinomial logit model containing individual covariates with the proportional odds model containing item attribute covariates.

$$P\left(\mathbf{y}_{nj} = \mathbf{r}|\alpha, \beta, \gamma, \pi\right) = \sum_{k=1}^{K} \pi_{nk} . P\left(\mathbf{y}_{nj} = \mathbf{r}|\alpha, \beta_k\right)$$

2   Implementation

In this work we concentrate on the more general second approach. The model is fitted using the EM algorithm and is implemented as a set of GLIM macros. The responses are converted to zero/one indicators that allow the use of the Poisson Likelihood in the model fit. The proportional odds model being a non-linear model can be accommodated using the OWN model facilities. The EM algorithm for fitting latent class models is equivalent to iterative fitting of a weighted GLM with posterior probabilities recalculated at each iteration. For the mixture model the EM algorithm is extended to include a step that refits the multinomial logit model.

3   Application

To illustrate the methodology a conjoint study of approximately 200 customers was conducted to investigate consumer car preferences. Five factors were identified as being key determinant attributes in the car market. The car attributes were brand, price and the number of doors and the individual characteristics were gender and age. The study compared 4 different price values, 4 brands and whether the car had 3 or 5 doors. We utilized a full profile method of collecting respondent evaluations. The design chosen

had two blocks of 16 cards each. The respondents were handed a set of 16 cards to compare with random assignment to block. The rating responses had seven categories where 1 corresponds to "worst" and 7 to "best".
The GLIM model formula for the utility model proposed by Green (2000) relates the utility of a product to its item attributes only.

$$(D + B * P\langle 2\rangle)\, .S$$

D is the number of doors attribute, B is the car brand, S is the segment and $P\langle 2\rangle$ is a quadratic function of price. This relationship allows a dual role for price; the negative price deterrent effect and a positive effect due to perceived quality. Models with four segments were used as they gave reasonable results in terms of choice behaviour. The deviance for this Proportional Odds Model was 9613. The relationship between worth and price was examined for each brand and segment through price profiles which characterise different customer behaviours. These include the strongly price sensitive customer who uses the price as a monetary constraint in choosing the item; those that use price as a signal of product quality and those with strong brand preferences.
In our first approach we included individual characteristics in the utility model to allow for individual differences in assessing the value of item characteristics.

$$(D * (A + G) + B * P\langle 2\rangle)\, .S$$

A and G are the respondents' age and gender. The deviance of this Proportional Odds Model was 9549. Although this model gave a significant reduction in deviance over the previous model it is very difficult to interpret. For example the parameter estimates show that the added worth of five-door cars increases more rapidly with age in segment 1 than other segments. Thus segment 1 will have more people who are either young and undervalue five-door cars or old and overvalue five-door cars. Such segments do not have a straightforward market interpretation and are not easy to target.
In our second approach we try to balance two competing goals; one is to obtain a model complex enough to provide a good fit and the other is to obtain a model that is simple to interpret. The Proportional Odds Model is as in Green (2000) and the multinomial logit model has model formula $A + G$. The deviance of this mixture model is 9560 which is comparable to the model presented in our first approach.
The Mixture model price profiles in figure 1 show the expected worth of each brand in the four fitted segments. Segment 1 represents consumers who have a moderate brand preference and are not strongly influenced by price. Respondents in segment 2 exhibit a strong reliance on price as a signal of quality but who hardly discriminate between the brands. People in segment 3 are differentiating between the brands and are price sensitive. Respondents in segment 4 have a strong brand preference and applying an
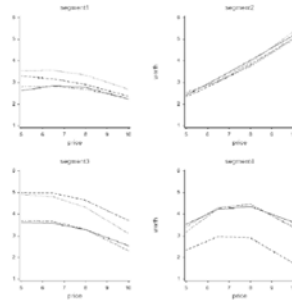
FIGURE 1.
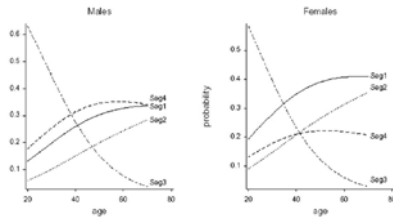
FIGURE 2.

"ideal price" as a signal that buying at very low prices could result in too low quality but see no bargain in buying at high prices.

Figure 2 shows the fitted model for segment membership probability as a function of age and gender. Segment 3 which is a cautious cost driven but brand selective group consists of a younger age group. Segments 1 and 2 consists of more females than males whereas segment 4 consists of more males than females for all ages. A marketer can more easily identify and target such segments.

4   Predicting preferences

Comparing the deviances of the two models is inadequate because the models are not nested. Standard diagnostic tools to check for outliers, influential data points and other model misspecifications cannot be used because the proportional odds model is a non linear and a non standard GLM. So a further task was included in the study in which each person was presented with four cards and choose the item that he/she preferred most. The purpose of this task was to observe how well our models predict peoples'

choice behaviour. For the extreme value distribution it is possible to derive the probability of preference from the predicted worth $\hat{W}_j$ The expected frequencies can hence be estimated by using the following result

$$P\left(\text{preference for } j^{th} \text{ item}\right) = \frac{\exp\left(\hat{W}_j\right)}{\exp\left(\hat{W}_1\right) + ... + \exp\left(\hat{W}_4\right)}$$

	Expected Frequency						Observed Frequency			
	Seg1	Seg2	Seg3	Seg4			Seg1	Seg2	Seg3	Seg4
S	24.8	7.66	7.41	23.5		S	16	4	7	20
P	6.80	2.15	31.9	12.3		P	13	8	19	11
O	3.29	1.22	11.0	4.36		O	9	5	9	8
F	14.1	22.0	11.7	1.84		F	11	16	27	3

The "observed" frequencies are defined by assigning individuals to segments with highest posterior probability and counting their first preferences. The expected frequencies are the totals of the predicted preference probabilities. Visual comparison of the observed and expected frequencies shows that the model is picking up the main features of individual preferences. It is eliciting that higher proportions of respondents in segments 1 and 4 prefer Subaru whereas a higher proportion of respondents in segment 2 like Fiat most. There is evidence that segment 3 is not a consistent predictor of individual preferences.

	Expected Frequency					Observed Frequency		
	Young	Old	Total			Young	Old	Total
S	25.19	38.16	63.35		S	18	29	47
P	32.16	21.02	53.18		P	32	19	51
O	11.63	8.24	19.87		O	15	16	31
F	24.02	25.59	49.61		F	28	29	57

The final two tables were produced to compare the number of preferred choices for different age groups using the Mixture Model. The model is correctly drawing out a higher proportion of old people rather than young ones who prefer Subaru and a higher proportion of young people rather than old ones who prefer Peugeot. It is rightly not eliciting any age bias for the other two brands. The Latent Class Model used in our first approach did similarly well. However the Mixture Model is effective in prediction of choice behaviour and leads to a segmentation model that has a clear and simple interpretation.

References

Green, M. (2000), Statistical Models for Conjoint Analysis, proceedings of the 15th IWMS, Bilbao, 216-222.

Models of double monotone dependence for two way contingency tables

Manuela Cazzaro[1] and Roberto Colombi[2]

[1] Università di Milano Bicocca - Piazza dell'Ateneo Nuovo - Milano - Italy; manuela.cazzaro@unimib.it
[2] Università di Bergamo - Viale Marconi - Dalmine - Italy; colombi@unibg.it

Abstract: To model the hypothesis of positive association between two categorical variables A and B a set of symmetric odds ratios defined on the joint probability function is usually subject to linear inequality constraints. In this paper two sets of asymmetric odds ratios defined respectively on the conditional distributions of A given B and on the conditional distributions of B given A are subject to linear inequality constraints.

1 Introduction

Let A and B be two ordered qualitative variables with r and c categories. Sometimes both the dependence of A from B and the dependence of B from A are of interest. For example, job satisfaction depends on insomnia and viceversa; the comfort of a waiting-room influences the perception of time of patients waiting for a medical examination and viceversa.
When both the dependence of A from B and the dependence of B from A are of interest we propose to constrain the $(r-1)(c-1)$ local-continuation odds ratios defined on the row conditional distributions of A given B and the $(r-1)(c-1)$ continuation-local odds ratios defined on the column conditional distributions of B given A. We prefer this approach to the usual one that constrains a set of $(r-1)(c-1)$ symmetric odds ratios, (*o.r.*), defined on the joint probabilities like the local (or global or continuation) odds ratios. If π_{ij} are the joint probabilities, the logarithms of the local-continuation odds ratios are defined on adjacent rows as follows:

$$\varphi_{ij} = \ln \frac{\pi_{ij} \cdot \sum_{m=j+1}^{c} \pi_{i+1m}}{\pi_{i+1j} \cdot \sum_{m=j+1}^{c} \pi_{im}}, \quad i = 1, 2, ..., r-1, \quad j = 1, 2, ..., c-1$$

and the logarithms of the continuation-local *o.r.* ψ_{ij} are analogously defined on adjacent columns of the contingency table. Alternatively the local-global and the global-local *o.r.*, which are similarly defined, can be used (for a survey of the various type of odds ratios see Douglas et al. (1990)).

2   Models and main results

A context that makes worthwhile imposing constraints on two types of odds ratios simultaneously is the case in which we want to test that both sets of conditional distributions are stochastically ordered. For example the hypothesis $\varphi_{ij} \geq 0$ and $\psi_{ij} \geq 0$ of double monotone dependence is equivalent to the hypothesis of uniform stochastic order of the row conditional distributions and of the column conditional distributions. A similar hypothesis of simple stochastic order is specified when the logarithms of the local-global and the global-local *o.r.* are assumed to be non-negative. For square tables the hypothesis of double monotone dependence will also be considered under the symmetry equality constraints: $\varphi_{ji} = \psi_{ij}$. Under this model the factors that multiply the continuation logits of the j-th row conditional distribution to obtain the corresponding logits in the next row are the same to the ones that give the continuation logits of the j-th column conditional distribution from the logits in the previous column. We are interested in testing the symmetry and double monotone dependence hypothesis against the symmetry alternative. First of all we show that only a subset of the $2(r-1)(c-1)$ inequalities $\varphi_{ij} \geq 0$ and $\psi_{ij} \geq 0$ is sufficient to express the condition of double monotone dependence. In fact note that the $(r-1)$ local continuation *o.r.* $\varphi_{i(c-1)}$ are *o.r.* of the local type and that $\varphi_{i(c-1)} \geq 0, i = 1, 2, ..., (r-1)$ implies that the conditional distribution of the c-th column is stochastically greater than the conditional distribution of the previous column according to the *likelihood ratio ordering*. Since the likelihood ratio stochastic ordering implies the *uniform ordering*, it follows that $\varphi_{i(c-1)} \geq 0$ implies $\psi_{i(c-1)} \geq 0$, for $i = 1, 2, ..., (r-1)$. Analogously we can state that $\psi_{(r-1)j} \geq 0$ implies $\varphi_{(r-1)j} \geq 0$ for $j = 1, 2, ..., (c-1)$. We should also note that $\varphi_{(r-1)(c-1)} = \psi_{(r-1)(c-1)}$. Therefore, in order to specify the double monotone dependence hypothesis, the following $(r-1)(c-1)+(r-2)(c-2)$ inequalities are sufficient:

$$\varphi_{ij} \geq 0, \psi_{ij} \geq 0, i = 1, 2, ...(r-2), j = 1, 2, ...(c-2),$$
$$\varphi_{i(c-1)} \geq 0, i = 1, 2, ...(r-1), \psi_{(r-1)j} \geq 0, j = 1, 2, ..., (c-2).$$

It is straightforward to verify, by means of counter examples, that the number of inequalities can not be further reduced.

It can be shown that for a $r \times c$ contingency table with r and c such that $[(r \geq 5) \cap (c \geq 5)] \cup [(r = 4) \cap (c \geq 7)] \cup [(r \geq 7) \cap (c = 4)]$ the number of inequalities needed to impose the double monotone dependence hypothesis is greater than the number of parameters $(rc - 1)$.

The double monotone dependence hypothesis can be imposed in square tables, when $r = c$, jointly with the symmetry hypothesis $\varphi_{ji} = \psi_{ij}$. In this case the number of inequalities specifying the double monotone dependence hypothesis can be further reduced. In fact $\varphi_{i(r-1)} \geq 0, i = 1, 2, ..., (r-1)$ implies not only $\psi_{i(r-1)} \geq 0, i = 1, 2, ..., (r-1)$, as in the general case

but also, for the symmetry hypothesis, $\psi_{(r-1)j} \geq 0, j = 1, 2, ..., (r-1)$. Furthermore, for the symmetry hypothesis, $\varphi_{ij} \geq 0, i = 1, 2, ...(r-2); j = 1, 2, ...(r-2)$ implies that $\psi_{ij} \geq 0, i = 1, 2, ...(r-2); j = 1, 2, ...(r-2)$. As a result the double monotone dependence and symmetry hypothesis is specified by the following $(r-1) + (r-2)^2$ inequalities:

$$\varphi_{ij} \geq 0, i = 1, ..., (r-2), j = 1, ..., (r-2); \varphi_{i(r-1)} \geq 0, i = 1, ..., (r-1).$$

It is worthwhile to note that these inequalities involve just a subset of the local-continuation o.r. so that they can be interpreted as linear constraints on the parameters of the parameterization of the joint probabilities based on the marginal distributions and the $(r-1)(c-1)$ local-continuation o.r. (see Colombi-Forcina (1999) for a discussion on this parameterization). Let $\boldsymbol{\theta}$ be the vector of the parameters of the saturated log-linear model of the joint probabilities π_{ij}. Moreover, let the symmetry equality constraints and the double monotone dependence inequality constraints be denoted by $\boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{0}$, $\boldsymbol{g}(\boldsymbol{\theta}) \geq \boldsymbol{0}$ (for the details on how these constraints can be written see Colombi-Forcina (2001)) and let $\boldsymbol{G}, \boldsymbol{H}$ be the jacobian matrices of $\boldsymbol{g}(\boldsymbol{\theta}), \boldsymbol{h}(\boldsymbol{\theta})$ at the point $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0$, which represents the unknown parameters vector under the hypothesis that all the inequality constraints are satisfied as equalities. Note that the previous constraints are non-linear in the parameters of the saturated log-linear model.
In the case of the hypothesis of double monotone dependence without symmetry, the number of inequality constraints is generally greater than $(rc-1)$, the dimension of the vector $\boldsymbol{\theta}$ of the log-linear parameters. In this case $\boldsymbol{G}$ has not full row rank, thus it is necessary to verify the Mangasarian-Fromovitz condition.
The Mangasarian-Fromovitz constraints qualification condition is satisfied at the point $\boldsymbol{\theta}_0$ if $C = \{\boldsymbol{d} : \boldsymbol{Gd} > \boldsymbol{0}, \boldsymbol{Hd} = \boldsymbol{0}\}$ is non empty. The just mentioned condition, easy to verify in our context, is relevant in Nonlinear Programming to establish necessary optimality criteria (Bazaraa et al. (1972), Mangasarian (1994)) and here, in the context of ordered restricted inference, is useful to obtain a reasonable asymptotic theory for the maximum likelihood estimators subject to inequality non linear constraints (Andrews (1999), Shapiro (1987)).

3   Examples

The proposed models will be illustrated through a data set concerning patients' satisfaction on various aspects of a medical service. In particular our data refer to a survey carried out in a national health service (NHS) trust of a northern italian city. These data have been collected by a telephone interview on about 2000 patients, concerning personal information and patients' satisfaction respect to waiting time, privacy protection and information received from doctors. In addition reservation of a specialist

examination, helpfulness of the staff, comforts of waiting-rooms, approach-ability of facilities, availability of suitable local transport have been anal-ysed. For example the dependence of the perception of the waiting-time from the comfort of the waiting-room and viceversa may be studied con-ditionally on explanatory variables such as age and education. In fact the models proposed in this work are used to analyse the data of Table 1, where medical service's users are asked to evaluate their satisfaction (*unsatisfied* (U), *satisfied* (S), *really satisfied* (RS)) regarding the waiting-room's com-fort (*COMFORT*) and the perception of the waiting-time (*TIME*) before a specialist examination is carried out. Moreover, patients are classified according to their age (*AGE*: ≤ 24, $25 - 54$, ≥ 55) and level of education (*EDUCATION*: *primary* (1), *secondary* (2), *high* (3)). The presence of co-variates implies that the considered hypotheses can be expressed with the previous established number of constraints for each subtable identified by the levels of the covariates.

TABLE 1. The NHS data.

EDUCATION		1			2			3		
COMFORT		U	S	RS	U	S	RS	U	S	RS
AGE	TIME									
	U	6	6	1	10	2	4	4	3	2
≤ 24	S	5	2	1	2	3	1	0	2	0
	RS	2	2	11	3	3	9	3	1	4
	U	37	11	13	48	20	13	31	12	3
$25 - 54$	S	20	17	9	23	24	17	12	8	5
	RS	11	18	49	25	22	66	7	14	38
	U	19	20	14	21	10	16	12	7	5
≥ 55	S	15	20	23	17	17	10	8	6	8
	RS	17	28	83	11	24	78	8	9	27

We test the double monotone dependence hypothesis between *COMFORT* and *TIME* with or without the symmetry hypothesis in each sub-table identified by the levels of the covariates, considering also the marginal con-tinuation logits of *COMFORT* and *TIME* as additive function of the effects of the covariates *AGE* and *EDUCATION*.

We report the likelihood ratio test statistic and the asymptotic simulated p-value (see Colombi-Forcina (2001) for the Monte Carlo method used to simulate the p-values) for the following models:

1: double monotone dependence, DMD, model, $G^2 = 8.83$, *p-value*=0.9966;

2: DMD and symmetry model, $G^2 = 2.43$, *p-value*=0.9974;

3: DMD and covariate additive effect model, $G^2 = 8.96$, *p-value*=0.9964;

4: DMD, symmetry and covariate additive effect model, $G^2 = 1.82$, *p-value*=0.9996. All the tested models show an excellent fit.

4   Conclusions

In summary the original topics and results of this work are:
1) it is shown that only a subset of the inequalities on the local-continuation and continuation-local odds ratios is necessary to model the hypothesis of double monotone dependence;
2) the double monotone dependence inequalities are non-linear in the log-linear parameters and generally their number is greater than the number $(rc - 1)$ of the parameters of the saturated log-linear model; however these inequalities satisfy the Mangasarian-Fromovitz condition so that the asymptotic distribution of the likelihood ratio statistics for testing the double monotone dependence hypothesis is easily obtained;
3) a data set concerning patients' satisfaction on a medical service is analysed in order to illustrate the usefulness of the new approach.

References

Andrews, D.W.K. (1999). Estimation when a parameter is on a boundary. *Econometrica*, **67**, 1341-1383.

Bazaraa, M. S., Goode, J. J., and Shetty C. M. (1972). Constraint Qualification Revisited. *Management Science*, **18**, 565-573.

Colombi, R., and Forcina, A. (1999). An instance of generalized log-linear models with inequality constraints: the continuation logit parametrization. In: *Proceedings of the 14th International Workshop on Statistical Modelling*. Edited by H. Friedl, Gratz.

Colombi, R., and Forcina, A. (2001). Marginal Regression Models for the Analysis of Positive Association of Ordinal Response Variables. *Biometrika*, **88**, 1007-1019.

Douglas, R., Fienberg, S. E., Lee, M. T., Sampson, A. R., Whitaker, L. R. (1990). Positive dependence concepts for ordinal contingency tables. In: *Topics in statistical dependence*, Block, H. W., Sampson, A. R. and Sanits, T. H. editors, Institute of Mathematical Statistics, Lecture Notes, Monograph series, **16**, 189-202, Haywar, California.

Mangasarian, O. L. (1994). *Nonlinear Programming*. SIAM, Philadelphia.

Shapiro, A. (1987). On differentiability of Metric Projection, 1: Boundary case. In: *Proceedings of the American Mathematical Society*. **99-1**, 123-128.

Non-parametric estimation of an intervention effect with staggered intervention times

Inês Sousa[1], Amanda Chetwynd[1] and Peter Diggle[1]

[1] Lancaster University, UK

Abstract: This talk is motivated by data from a longitudinal trial comparing the progress of patients randomised between two treatment groups, one with and one without surgical intervention, in which the time of the surgical intervention varies between patients. Our aim is to obtain non-parametric estimators of the longitudinal mean response in the non-surgical arm, and the surgical intervention effect.

Keywords: cubic smoothing spline; back-fitting algorithm; cross-validation; longitudinal data analysis; non-parametric estimation

1 Introduction

This work is motivated by data from a longitudinal trial on Lung Emphysema. It compares the progress of patients randomised between two treatment groups, one with and one without surgical intervention. Surgical intervention time is subject-specific. The response variable is forced expiratory volume in one second (FEV_1). Our aim is to obtain non-parametric estimators of the longitudinal mean response in the non-surgical arm, and the surgical intervention effect.

2 The Model

Standard methods of exploratory data analysis are not well suited to this specific data set, because of the patient-specific surgical intervention times. We propose a method for exploratory data analysis using non-parametric spline-smoothing.

Suppose subject i provides a sequence of responses y_{ij} at times t_{ij}, and the time of surgical intervention, if any, is s_i. Write

$$y_{ij} = \mu_i(t_{ij}) + \varepsilon_{ij} \tag{1}$$

where the errors ε_{ij} are correlated within subjects.

We assume that,

$$\mu_i(t_{ij}) = \begin{cases} \mu_0(t_{ij}) & : \quad t_{ij} < s_i \\ \mu_0(t_{ij}) + \delta(t_{ij} - s_i) & : \quad t_{ij} \geq s_i. \end{cases} \tag{2}$$

In (2), the function $\mu_0(\cdot)$ is interpreted as the mean response under the standard treatment, i.e. without surgical intervention, whilst the function $\delta(\cdot)$ is the mean longitudinal effect of surgical intervention, as a function of time since surgery. Our aim is to obtain smooth, non-parametric estimates of $\mu_0(\cdot)$ and $\delta(\cdot)$.

3   Estimation and Inference

Our roughness penalty estimation method is penalized sum of squares criterion, with a term for each of the functions to be estimated. The criterion is then defined for any two twice-differentiable functions, with assumed smoothing parameters $\lambda_1, \lambda_2 > 0$, as

$$S(\mu_0, \delta) \quad = \quad \sum_{i=1}^{m} \sum_{j=1}^{n_i} \{y_{ij} - \mu_i(t_{ij}, s_i)\}^2 +$$

$$+ \lambda_1 \int \mu_0''^2(x)dx + \lambda_2 \int \delta''^2(x)dx \qquad (3)$$

Where $\mu_i(t_{ij})$ is given by (2). We prove that the functions $\hat{\mu}_0(\cdot)$ and $\hat{\delta}(\cdot)$, which minimise (3), are natural cubic smoothing splines. For given values of λ_1 and λ_2 we then obtain the estimates $\hat{\mu}_0(\cdot)$ and $\hat{\delta}(\cdot)$ by a back-fitting algorithm (Hastie, 1990).

To choose the values of λ_1 and λ_2 we use a cross-validation criterion defined as in Rice (1991), which allows for the correlation between repeated measurements on the same subject by deleting all measurements on one subject at a time, rather than one measurement at a time.

To obtain interval estimates of $\mu_0(\cdot)$ and $\delta(\cdot)$, we use a Monte Carlo method as follows. Using the estimates $\hat{\mu}_0(\cdot)$ and $\hat{\delta}(\cdot)$ we construct residuals, $r_{ij} = y_{ij} - \hat{\mu}_i(t_{ij})$. We then compute the empirical variogram of the r_{ij} (Diggle, 2002) and use non-linear ordinary least squares to fit a parametric error model including terms for a random subject-specific intercept, serially correlated random variation over time within each subject, and measurement error. Finally, we simulate 300 data-sets from the resulting model, re-estimate the functions $\mu_0(\cdot)$ and $\delta(\cdot)$ from the simulated data-sets and compute pointwise quantiles of the re-estimates at each time-point.

4   Results

We illustrate our methodology in the Lung Emphysema data set. Figure 1 shows the estimate functions $\hat{\mu}(.)$ and $\hat{\delta}(.)$ together with their envelop intervals obtained using the new methodology.
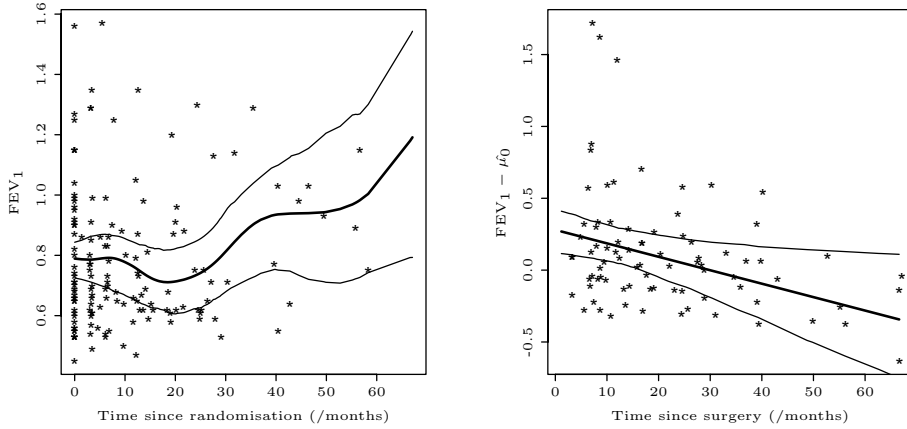
FIGURE 1. $\hat{\mu}_0$ and $\hat{\delta}$ with respective envelop intervals

References

Diggle, P.J., Heagerty, P.J., Liang, K-Y., and Zeger, S.L. (2002). *Analysis of Longitudinal Data*. Oxford University Press.

Hastie, T.J., and Tibshirani, R.J. (1990). *Generalized Additive Models*. Number 43 in Monographs on Statistics and Applied Probability. Chapman & Hall

Rice, A., and Silverman, B.W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Series B*, **53**, 233-243.

Exploratory analysis of epidemiological time series by means of transfer function models

Monica Chiogna[1] and Carlo Gaetan[2]

[1] Dipartimento di Scienze Statistiche - Università di Padova
[2] Dipartimento di Statistica - Università Ca' Foscari - Venezia

Abstract: The objective of this paper is to explore the potential of the transfer function methodology for exploratory analysis of data in multi-site epidemiological time series studies. The ideas are illustrated by analysing data on the relationship between daily non accidental deaths and air pollution in the 20 US largest cities.

1 Introduction

Time series studies are specially suitable in epidemiology for evaluating short-term effects on human health of time-varying exposures to air pollution. The methodology most frequently adopted relies on regression, i.e. disease or death occurrences are related to the suspected risk factors by regressing counts aggregated over geographical units on aggregated covariate summaries. Standard regression methods used initially have been nowadays almost fully replaced by semi-parametric approaches, such as semi-parametric generalized additive models (Hastie and Tibshirani, 1990).
Recent multi-site studies (Dominici *et al.*, 2000; Biggeri *et al.*, 2001; Atkinson *et al.*, 2001) have shown that combination of data from disparate sources provides additional statistical power to the analysis, that it is not available in single site analyses. Clearly, construction of a model to be used in the meta-analysis becomes rapidly more complicated as the number of cities increases.
In this work, we wish to investigate the potential of transfer function analysis in providing an affordable computational framework to allow exploratory analysis of the relations among the time series used in the models. In fact, when dealing with many sources of data, an exploratory analysis on which to base model construction becomes rapidly unaffordable. Our idea is to use indications coming from a data-driven model selection to highlight the common features across sites. We illustrate these ideas by analysing data on the relationship between daily non accidental deaths and air pollution in the 20 largest US cities.

2   Moving to transfer function models

Let $C(t)$ be the dependent variable, like, for example, the daily number of deaths. The independent variables are of three types: covariates representing temporal patterns, meteorological variables, and air pollution concentrations. Standard techniques for building the models are based Poisson regression. The common model construction strategy develops in three steps: (1) adjusting for temporal confounding; (2) adjusting for meteorological confounding; (3) inserting pollutant(s).

When $C(t)$ is high, often the response is transformed to bring the models back to regression settings. Let $Y(t)$ indicate the transformed response. The final model, with a single pollutant $Z(t)$, takes the following form:

$$Y(t) = T(t) + M(t) + \beta Z(t) + n(t)$$

where $T(t)$ and $M(t)$ are suitable functions representing temporal trends and the effects of meteorology and $n(t)$ is a noise term.

To show how transfer function models can be used as modelling strategy in this context, let us consider first the problem of adjusting for temporal confounding. At this stage, the model to be built is of type

$$Y(t) = T(t) + n(t).$$

Usually, $T(t)$ is modelled nonparametrically. A discrete time analog of one such model with a continuous-time cubic spline can be written as an ARIMA(0,2,1) process observed with error:

$$Y(t) = T(t) + \lambda \eta(t), \qquad (1 - B)^2 T(t) = (1 + \theta B)\xi(t),$$

where $(\eta(t), \xi(t)) \sim (0, \sigma^2)$ and B is the lag operator, i.e. $BY(t) = Y(t-1)$. It can be shown (Hyndmann *et al.*, 2004) that this is equivalent to an ARIMA(0,2,2) model with some restriction on the parameters. In modelling temporal confounding, the problem is to capture lagged effects. This is done by using (often linear) functions of past values, like, for example, distributed lag models.

It is easy to see that all the components which enter the final model can be assembled in a structure of type:

$$Y(t) = \sum_{i=1}^{I} \frac{\omega_i(B)}{\delta_i(B)} X_i(t) + \frac{\theta(B)}{(1 - B)^d (1 - B^s)^D} e(t), \qquad (1)$$

where $\{X_i(t)\}$, $i = 1, \ldots, I$, are the covariates of interest, $\{e(t)\}$, is a zero-mean stationary process independent of the covariates, $\omega_i(B) = \omega_{i0} - \omega_{i1} - \ldots - \omega_{ir_i}$, $\delta_i(B) = 1 - \delta_{i1} - \ldots - \delta_{is_i}$, $\theta_i(B) = 1 - \theta_{i1} - \ldots - \theta_{iq}$, are polynomials in the lag operator B, with degrees r_i, u_i, q respectively. Setting 1 defines a transfer function (TF) model. In equation (1), the roots of the polynomials $\delta_i(z)$, $i = 1, \ldots, I$, are supposed to be outside of the unit circle.

3   Identification of TF models by an iterative stepwise algorithm

An appropriate model can be selected by searching within the space of models defined by equation (1), after having selected the covariates and the degrees of the polynomials. The vector α of unknown coefficients of (1) can be estimated using a prediction error method, by minimizing

$$s(\alpha) = \frac{\sum_{t=\tilde{t}+1}^{T} e(t;\alpha)^2}{\tilde{T}}. \tag{2}$$

As well known, the estimate $\hat{\alpha}$ cannot be computed analytically. To solve this nonlinear least-squares problem we use a Gauss-Newton type algorithm

$$\hat{\alpha}_{n+1} = \hat{\alpha}_n - \lambda_n \left[\sum_{t=1}^{T} J(t;\hat{\alpha}_n)J(t;\hat{\alpha}_n)^T \right]^{-1} \sum_{t=1}^{T} J(t;\hat{\alpha}_n)e(t;\hat{\alpha}_n),$$

where $0 < \lambda_n \leq 1$ and $J(t;\hat{\alpha}_n)$ is the Jacobian vector $\partial \hat{Y}(t;\alpha)/\partial \alpha$. Note that $\hat{\alpha}_{n+1}$ is the least square solution of $J(t;\hat{\alpha}_n)^T \hat{\alpha}_n - \lambda_n e(t;\hat{\alpha}_n) = J(t;\hat{\alpha}_n)^T \alpha$, $t = 1, \ldots, T$.

This remark allows us to couple the estimation process with the selection of the lag structure. More precisely, we propose this identification procedure: (1) choose a starting value $\hat{\alpha}_0$; (2) solve the least square problem and get α_{n+1}^*; (3) apply a backward stepwise selection procedure to α_{n+1}^* according to a selection criterion such as $BIC = \log s(\hat{\alpha}) + m\frac{\log \tilde{T}}{\tilde{T}}$, and obtain $\hat{\alpha}_{n+1}$; (4) set $n = n + 1$ and return to 2 until a converge criterion is met.

4   Modelling in practice

The data that we consider come from the National Morbidity, Mortality and Air Pollution Study (NMMAPS, Dominici *et al.*, 2000), to which we refer for further details about sources of the data. Data are available at the URL `http://ihapss.biostat.jhsph.edu/data/`. In our analysis, we will explore the association between daily changes in the concentration of carbon monoxide (CO) and daily number of deaths in the 20 US largest cities, for which NMMAPS reports positive significant effects of CO at the usual lags (0,1,2). In our example $\{X_1(t)\}$, $\{X_2(t)\}$, $\{X_3(t)\}$, are the pollutant, temperature and dew point time series, respectively. As the mean number of counts is sufficiently high, we can safely consider the transformation $Y(t) = \sqrt{C(t)}$ and move to linear models. This allows us to connect to the transfer function methodology.

To offer our model section procedure a great deal of flexibility, we chose the following model setting:

$$Y(t) = \sum_{i=1}^{3} \omega_i(B)X_i(t) + \frac{\theta(B)}{(1-B)(1-B^7)}e(t),$$

where $r_i = 3$, $i = 1, \ldots, 3$ and $q = 7$. This formulation allows to take into account short term seasonal patterns and long term trends. Moreover, it allows to incorporate lagged values of the inputs. Based on evidence from the literature, we considered that the first three lags were sufficient to catch delayed effects of the covariates. Note that, despite the relative simplicity of this model formulation, cardinality of the model space is around 2.1×10^6. At the end of the model selection, in 8 cities found the search strategy a significant effect of CO. In all the 20 cities, the selection procedure adopted first order differences for the input and the output series.

5   Results

To perform the meta-analysis task, we adopted the strategy of fitting the same common model to all the cities and to combine evidence resulting from the model fitting. Based on the output from the automated model selection procedure, we decided to fit the following common model:

$$Y(t) = \omega_{11} X_1(t-1) + \omega_{23} X_2(t-3) + \omega_{32} X_3(t-2) + \frac{1 - \theta B}{1 - B} e(t).$$

The common model allowed to detect a significant effect in 11 cities. In the meta-analysis, the estimates for CO for each city were combined using fixed and random effects models (Normand, 1999). As expected, the confidence intervals were wider under the random effects model, and narrower under the fixed effects model. Nevertheless, differences in point estimates were negligible. City-specific and pooled estimates for the random effects model are represented graphically in Figure 1. A geographical gradient in value for the effect is visible, with Seattle and Minneapolis distinguishing from the remaining cities. Estimates are significant in Southern California and in the Southwest, become not significant moving to the Southeast, and return significant moving to the Northeast and industrial Midwest. This agrees with the effects found for PM_{10} from NMMAPS.

References

Atkinson, R.W., Anderson , H.R., Sunyer, J., Ayre , J., Baccini, M., Vonk, J.M., Boumghar, A., Forastiere, F., Forsberg, B., Touloumi , G., Schwartz, J. and Katsouyanni, K. (2001). Acute effects of particulate air pollution on respiratory admissions: results from APHEA 2 project. Air Pollution and Health: a European Approach. *Am J Respir Crit Care Med,* **15**, 164(10 Pt 1), 1860–1866.
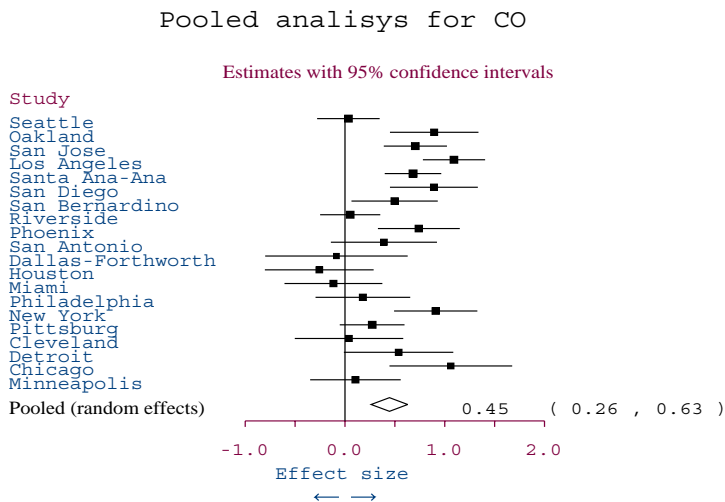
FIGURE 1. Results of the pooled analysis under the random effects model (coefficients are multiplied by 10^4).

Biggeri, A., Bellini, P. and Terracini, B. (2001). Meta-analysis of the Italian studies on short-term effects of air pollution, *Epidemiol. Prev.*, **25**(2 Suppl), 1-71. In Italian.

Dominici, F., Zeger, S.L. and Samet, J.M. (2000). Combining Evidence on Air Pollution and Daily Mortality from the Largest 20 US cities: A Hierarchical Modeling Strategy (with discussion). *Journal of the Royal Statistical Society, Series A*, **163**, 263-302.

Hastie, T.J., Tibshirani, R.J. (1990). *Generalized additive models*. Chapman and Hall, New York.

Hyndman, R.J., King, M.L., Pitrun, I., Baki Billah, B. (2004). Local linear forecasts using cubic smoothing splines. *Australian and New Zealand Journal of Statistics*, to appear.

Normand (1999). Meta analysis: Formulating, evaluating, synthesizing, and reporting. *Statistics in Medicine*, **18**, 321–359.

Efficient smoothing of d-dimensional arrays

Iain Currie[1], Maria Durbán[2] and Paul Eilers[3]

[1] Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, EH14 4AS, Scotland
Email: `I.D.Currie@ma.hw.ac.uk`
[2] Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, Madrid, Spain
[3] Department of Medical Statistics, Leiden University Medical Center, 2300 RC Leiden, The Netherlands

Abstract: Data on multidimensional arrays are wide-spread and modelling can easily present storage and computational difficulties, even with modern computers. We present a class of regression models and a computational procedure designed specifically for such data. These models possess some remarkable storage and computational properties which lead to savings of orders of magnitude in both storage and speed over conventional methods. We call this methodology array regression. We illustrate our procedure with the analysis of a large set of count data on deaths from respiratory disease indexed by age of death, year of death and month of death.

Keywords: Arrays; GLM; Kronecker product; P-splines; Smoothing.

1   Array regression: what is it?

In this paper we analyse a set of count data indexed by age of death (1 to 105), year of death (1959 to 1998) and month of death (1 to 12). The 50400 data points are arranged in a 3-dimensional array whose sides have length 105, 40 and 12. Suppose we summarize the data by a coarser array whose sides have size 10, 5 and 3, say; the summary array will have 150 cells with entries obtained by some kind of local averaging. The idea is to use this array as parameters in a regression model. The estimation of the regression coefficients could be done using the usual regression approach of "flattening" both the data array and the coefficient array. However this approach fails to make use of the structure of the data and leads directly to the "curse of dimensionality". In array regression we avoid computation of the full "flattened" regression matrix and instead reduce the fitting of the model to a sequence of operations whose storage and computational load is determined by the lengths of the sides of both the data and coefficient arrays.

One important setting for these ideas is multidimensional smoothing which we consider in the penalized generalized linear model (PGLM) framework.

Eilers and Marx (1996) use penalized B-splines to smooth 1-dimensional data and their algorithm

$$(\boldsymbol{B}'\tilde{\boldsymbol{W}}_\delta\boldsymbol{B} + \boldsymbol{P})\hat{\boldsymbol{\theta}} = \boldsymbol{B}'\tilde{\boldsymbol{W}}_\delta\boldsymbol{B}\tilde{\boldsymbol{\theta}} + \boldsymbol{B}'(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}) \tag{1}$$

is a generalization of the standard scoring algorithm for a GLM. We note that $\boldsymbol{B}$ is a banded matrix with $\boldsymbol{B}\mathbf{1} = \mathbf{1}$ and $\boldsymbol{B} \geq \mathbf{0}$, so B-splines provide a suitable basis for local averaging. The ingredients of a PGLM are thus: the vectors of observations $\boldsymbol{y}$, means $\boldsymbol{\mu}$, offsets $\boldsymbol{o}$ (if any), and regression coefficients $\boldsymbol{\theta}$, the diagonal matrix of weights, $\boldsymbol{W}_\delta$, the regression matrix, $\boldsymbol{B}$, and the penalty matrix, $\boldsymbol{P}$. We can represent this schematically as $\{\boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{o}, \boldsymbol{W}_\delta, \boldsymbol{B}, \boldsymbol{P}, \boldsymbol{\theta}\}$. The computational demands in (1) are of two kinds: linear functions $\boldsymbol{B}\boldsymbol{\theta}$ (the linear predictor) and $\boldsymbol{B}'(\boldsymbol{y} - \boldsymbol{\mu})$, and inner products $\boldsymbol{B}'\boldsymbol{W}_\delta\boldsymbol{B}$. In contrast the scheme in array regression for data in a d-dimensional array has the form $\{\boldsymbol{Y}, \boldsymbol{M}, \boldsymbol{O}, \boldsymbol{W}, \boldsymbol{B}_1, \boldsymbol{B}_2, \ldots, \boldsymbol{B}_d, \boldsymbol{P}, \boldsymbol{\Theta}\}$ where $\boldsymbol{Y}$, $\boldsymbol{M}$, $\boldsymbol{O}$, $\boldsymbol{W}$ and $\boldsymbol{\Theta}$ are d-dimensional arrays, $\boldsymbol{B}_1$, $\boldsymbol{B}_2 \ldots \boldsymbol{B}_d$ is a set of 1-dimensional B-spline bases defined on each variable in turn, and $\boldsymbol{P}$ is the penalty matrix. The d-dimensional basis $\boldsymbol{B}$ is the Kronecker product of the 1-dimensional bases. The computational demands are again to compute the linear functions and inner products in (1) but these demands are met with a new set of tools.

2   Array regression: how to perform it

Currie, Durban and Eilers (2003) used a PGLM to smooth a 2-dimensional mortality table indexed by year of death and age of death. They argued that an appropriate regression matrix was $\boldsymbol{B}_y \otimes \boldsymbol{B}_a$ where $\boldsymbol{B}_y$ and $\boldsymbol{B}_a$ were regression matrices of B-splines on the marginal variables year and age. We generalize this and suppose that the data are arranged in a d-dimensional array $\boldsymbol{Y}$, $n_1 \times \ldots \times n_d$, and use

$$\boldsymbol{B} = \boldsymbol{B}_d \otimes \ldots \otimes \boldsymbol{B}_1 \tag{2}$$

as regression matrix; here $\otimes$ is the Kronecker product and $\boldsymbol{B}_i$ is $n_i \times c_i$, $i = 1, \ldots, d$. (This representation assumes that the array is stored with the first dimension varying fastest, the second dimension varying next and so on, as in Splus, for example.) The regression matrix $\boldsymbol{B}$ inherits the properties $\boldsymbol{B}\mathbf{1} = \mathbf{1}$ and $\boldsymbol{B} \geq \mathbf{0}$ so provides a suitable basis for local averaging in d-dimensions; the regression coefficients $\boldsymbol{\theta}$ are regarded as a $c_1 \times \ldots \times c_d$ dimensional array $\boldsymbol{\Theta}$. However, $\boldsymbol{B}$ can quickly become very large and the standard approach of flattening the data and proceeding with the usual regression algorithm is either very slow or simply not available. We develop a new algorithm which takes advantage of the structure of both the data and the regression model. We make four definitions.

Definition 1: The *row tensor* of a matrix $\boldsymbol{X}$ with c columns is defined as

$$G(\boldsymbol{X}) = (\boldsymbol{X} \otimes \mathbf{1}') * (\mathbf{1}' \otimes \boldsymbol{X}) \tag{3}$$

where $\mathbf{1}$ is a vector of 1's of length c and $*$ denotes element by element multiplication.

Definition 2: The *H-transform* of the d-dimensional array $\boldsymbol{A}$ of size $c_1 \times c_2 \ldots \times c_d$ by the matrix $\boldsymbol{X}$ of size $r \times c_1$ is denoted $H(\boldsymbol{X}, \boldsymbol{A})$ and defined as follows: let $\boldsymbol{A}^*$ be the $c_1 \times c_2 c_3 \ldots c_d$ matrix obtained by flattening dimensions 2 to d of $\boldsymbol{A}$; form the matrix product $\boldsymbol{X}\boldsymbol{A}^*$ of size $r \times c_2 c_3 \ldots c_d$; then $H(\boldsymbol{X}, \boldsymbol{A})$ is the d-dimensional array of size $r \times c_2 \ldots \times c_d$ obtained from $\boldsymbol{X}\boldsymbol{A}^*$ by reinstating dimensions 2 to d of $\boldsymbol{A}$.

Definition 3: We define the *rotation* of the d-dimensional array $\boldsymbol{A}$ of size $c_1 \times c_2 \ldots \times c_d$ to be the d-dimensional array $R(\boldsymbol{A})$ of size $c_2 \times c_3 \ldots \times c_d \times c_1$ obtained by permuting the indices of $\boldsymbol{A}$.

Definition 4: We define the *rotated H-transform* of the array $\boldsymbol{A}$ by the matrix $\boldsymbol{X}$ by $\rho(\boldsymbol{X}, \boldsymbol{A}) = R(H(\boldsymbol{X}, \boldsymbol{A}))$.

The tools for the computation of the linear functions $\boldsymbol{B}\boldsymbol{a}$ and $\boldsymbol{B}'(\boldsymbol{y} - \boldsymbol{\mu})$, and the inner product $\boldsymbol{B}'\boldsymbol{W}_\delta\boldsymbol{B}$ can now be stated:

Linear function: The elements of $\boldsymbol{B}\boldsymbol{\theta}$ (and similarly for $\boldsymbol{B}'(\boldsymbol{y} - \boldsymbol{\mu})$) are given by the d-dimensional array

$$\rho(\boldsymbol{B}_d, \ldots, \rho(\boldsymbol{B}_2, \rho(\boldsymbol{B}_1, \boldsymbol{\Theta}))\ldots). \tag{4}$$

Inner product: The elements of the inner product $\boldsymbol{B}'\boldsymbol{W}_\delta\boldsymbol{B}$ are given by the d-dimensional array

$$\rho(G(\boldsymbol{B}_d)', \ldots, \rho(G(\boldsymbol{B}_2)', \rho(G(\boldsymbol{B}_1)', \boldsymbol{W}))\ldots). \tag{5}$$

The vectors $\boldsymbol{B}\boldsymbol{\theta}$ and $\boldsymbol{B}'(\boldsymbol{y} - \boldsymbol{\mu})$, and the matrix $\boldsymbol{B}'\boldsymbol{W}_\delta\boldsymbol{B}$ are obtained by rearrangement and re-dimensioning of (4) and (5); we omit details of this in the present paper. The important feature of (4) and (5) is that they avoid storage of the full regression matrix $\boldsymbol{B}$ and require far fewer multiplications. It remains to define the penalty matrix $\boldsymbol{P}$. The expression in 3-dimensions indicates the general formula. We penalize each dimension in turn, i.e., we place penalties on the rows, columns, etc of the array. We find

$$\boldsymbol{P} = \lambda_1 \boldsymbol{I}_{c_3} \otimes \boldsymbol{I}_{c_2} \otimes \boldsymbol{D}_1'\boldsymbol{D}_1 + \lambda_2 \boldsymbol{I}_{c_3} \otimes \boldsymbol{D}_2'\boldsymbol{D}_2 \otimes \boldsymbol{I}_{c_1} + \lambda_3 \boldsymbol{D}_3'\boldsymbol{D}_3 \otimes \boldsymbol{I}_{c_2} \otimes \boldsymbol{I}_{c_1} \tag{6}$$

where $\boldsymbol{D}_1$, $\boldsymbol{D}_2$ and $\boldsymbol{D}_3$ are difference matrices.

3   Array regression: an example

We illustrate our method with some data on the number of deaths from respiratory disease. The data array $\boldsymbol{Y} = Y[i, j, k]$ is indexed by age of death, $i = 1, \ldots, 105$, year of death, $j = 1, \ldots, 40$ (1959 to 1998) and month of death, $k = 1, \ldots, 12$. Thus $\boldsymbol{Y}$ has 50400 points arranged in a $105 \times 40 \times 12$ array. We assume that the number of deaths $Y[i, j, k]$ can be modelled by a PGLM with Poisson error and log link; the log of the number of days in a month is used as an offset. The regression matrix is defined
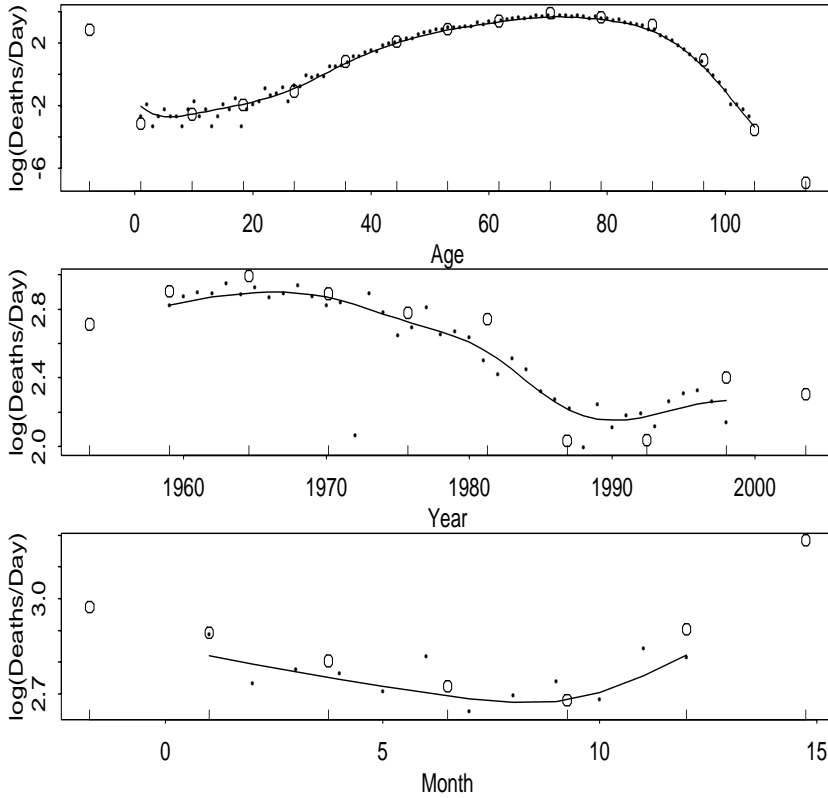
FIGURE 1. Observed and smoothed numbers of log(deaths/day) by age, year and month. Regression coefficients $\hat{\boldsymbol{\Theta}}[i,j,k]$ are plotted ∘ against knot position. Top panel: January, 1959, $\hat{\boldsymbol{\Theta}}[i,2,2], i = 1, \ldots, 15$; middle panel: age 53, January, $\hat{\boldsymbol{\Theta}}[8,j,2], j = 1, \ldots, 10$; bottom panel: age 53, 1959, $\hat{\boldsymbol{\Theta}}[8,2,k], k = 1, \ldots, 7$.

via the marginal regression matrices of B-splines for age, year and month. We choose knots as follows: at 1 and 105 with 11 internal knots for age, at 1 and 40 with 6 internal knots for year, and at 1 and 12 with 3 internal knots for month. With cubic B-splines this gives $\boldsymbol{B}_1$, 105×15, $\boldsymbol{B}_2$, 40×10 and $\boldsymbol{B}_3$, 12×7. The regression matrix has 1050 parameters arranged in a $15 \times 10 \times 7$ array. This is a large regression problem: the regression matrix $\boldsymbol{B}$ alone has over 5×10^7 elements. The parameters are estimated using second order penalties and the Bayesian Information Criterion (BIC). The fitted model has effective degrees of freedom of 305.

Figure 1 gives some idea of how the numbers of deaths vary with age, year and month. The plots also show how the coefficient array $\boldsymbol{\Theta}$ approximates the data array $\boldsymbol{Y}$ (on the scale of the linear predictor). A smoothed value

at age i, year j and month k is a weighted average of elements in the coefficient array where the weights are given by the Kronecker product of rows of the marginal regression matrices, $\boldsymbol{B}_3[k,] \otimes \boldsymbol{B}_2[j,] \otimes \boldsymbol{B}_1[i,]$. The non-zero weights apply to a sub-array of coefficients (generally $4 \times 4 \times 4$ with cubic B-splines) in the vicinity of $Y[i, j, k]$.

We conclude with some remarks on the performance of our approach. The most demanding component of (1) is the calculation of $\boldsymbol{B}'\boldsymbol{W}_\delta\boldsymbol{B}$; this requires the multiplication of two large matrices. Absolute timings are machine dependent so the ratio of the speeds of the two methods is of greater interest. Table 1 shows that the larger the coefficient array the greater the gain with array regression over standard regression. For a $9 \times 9 \times 9$ coefficient array we were unable to store the full regression matrix $\boldsymbol{B}$.

TABLE 1. Times (seconds) to calculate $\boldsymbol{B}'\boldsymbol{W}_\delta\boldsymbol{B}$

Array size	npar	Standard regression	Array regression	Ratio
$6 \times 6 \times 6$	216	20	1	20:1
$7 \times 7 \times 7$	343	200	2	100:1
$8 \times 8 \times 8$	512	2000	4	500:1
$9 \times 9 \times 9$	729	–	20	–

4   Array regression: conclusions

Array regression is a fast, low storage method designed for smoothing multidimensional arrays. The method uses penalized regression to smooth data using a local averaging algorithm. The important feature of our method is that the local averaging is performed sequentially, dimension by dimension, thus avoiding the full impact of the "curse of dimensionality".

References

Currie, I., Durban, M. & Eilers, P. (2003). Smoothing and forecasting mortality rates. Submitted to *Statistical Modelling*.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statist. Sci.* **11**, 89-121.

Assessment of Variance Components in Elliptical Linear Mixed Models

Carine Savalli[1], Gilberto A. Paula[1] and Francisco José A. Cysneiros[2]

[1] Instituto de Matemática e Estatística, USP - Caixa Postal 66281 (Ag. Cidade de São Paulo), 05311-970, São Paulo - SP - Brazil, e-mail: carinesr@ime.usp.br and giapaula@ime.usp.br

[2] Departamento de Estatística, Universidade Federal de Pernambuco, Recife - PE - Brazil, e-mail: cysneiros@de.ufpe.br

Abstract: This paper's aim is to discuss the problem of testing variance components in elliptical linear mixed models. The elliptical class includes all symmetrical continuous distributions, such as normal, Student-t, Pearson VII, exponential power and logistic, among others. A score-type test for one-sided alternatives is applied and an illustrative example for which a Student-t distribution is assumed for the responses and random effects is presented. The results are compared with the ones from the normal mixed model.

Keywords: Hypothesis testing; Variance Components; Elliptical distributions; Robust models; Score tests; One-sided alternatives.

1 Introduction

The importance of linear mixed models for analyzing repeated measures data with continuous normal responses is undeniable. A general hierarchical structure proposed by Laird and Ware (1982) assumes that

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \ i = 1, \ldots, n, \tag{1}$$

where $\mathbf{y}_i$ is an m_i-dimensional random vector of observed responses from the ith cluster, $\mathbf{X}_i$ is an $m_i \times p$ matrix which contains values of p explanatory variables, $\boldsymbol{\beta}$ is the fixed parameter vector, $\mathbf{Z}_i$ is an $m_i \times q$ design matrix of random effects $\mathbf{b}_i$ and $\boldsymbol{\epsilon}_i$ is an m_i-dimensional vector of within-cluster errors. It is usual to assume $\mathbf{b}_i \sim \mathrm{N}_q(\mathbf{0}, \mathbf{D})$ and $\boldsymbol{\epsilon}_i \sim \mathrm{N}_{m_i}(\mathbf{0}, \sigma^2\mathbf{I}_{m_i})$. However, due to lack of robustness of normal models against extreme observations, a general class of elliptical models can be preferred to overcome this problem. The elliptical class includes all symmetrical continuous distributions, such as normal, Student-t, Pearson VII, exponential power and logistic, among others, and their properties are described in Fang, Kotz and Ng, (1990). To deal with extreme observations, for example, instead of assuming normality

for $\mathbf{b}_i$ and $\boldsymbol{\epsilon}_i$ we can assume that $(\mathbf{b}_i^T, \boldsymbol{\epsilon}_i^T)^T$ follows a Student-t distribution of mean zero and dispersion matrix $\mathbf{V}_i = \text{diag}\{\mathbf{D}, \sigma^2 \mathbf{I}_{m_i}\}$, namely, $(\mathbf{b}_i^T, \boldsymbol{\epsilon}_i^T)^T \sim \text{El}(\mathbf{0}, \mathbf{V}_i)$. It means that $\mathbf{b}_i$ and $\boldsymbol{\epsilon}_i$ are uncorrelated but not necessarily independent (unless for the normal case). Thus, we can express

$$\begin{bmatrix} \mathbf{y}_i \\ \mathbf{b}_i \end{bmatrix} \sim \text{El} \left\{ \begin{pmatrix} \mathbf{X}_i \boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix}; \begin{bmatrix} \sigma^2 \mathbf{I}_{m_i} + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T & \mathbf{Z}_i \mathbf{D} \\ \mathbf{D} \mathbf{Z}_i^T & \mathbf{D} \end{bmatrix} \right\}, \; i = 1, \ldots, n. \quad (2)$$

2   Marginal Elliptical Model

Similar to normal mixed models, inferences in elliptical mixed models may be based on the marginal distribution of $\mathbf{y}_i$, which takes the form

$$\mathbf{y}_i \sim \text{El}(\mathbf{X}_i \boldsymbol{\beta}; \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \sigma^2 \mathbf{I}_{m_i}). \quad (3)$$

The density function of $\mathbf{y}_i$ is given by

$$f(\mathbf{y}_i) = |\boldsymbol{\Sigma}_i|^{-1/2} g(u_i), \; i = 1, \ldots, n, \quad (4)$$

where $u_i = (\mathbf{y}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)$ with $\boldsymbol{\Sigma}_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \sigma^2 \mathbf{I}_{m_i}$, $g(.) : \mathbb{R} \to [0, \infty]$ so that $\int_0^\infty u^{m_i/2-1} g(u) du < \infty$ called density generator (see, for example, Fang, Kotz and Ng, 1990), $\boldsymbol{\mu}_i = \mathbf{X}_i \boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_i$ is proportional to the variance-covariance matrix of $\mathbf{y}_i$. For simplicity we will assume $\mathbf{D} = \text{diag}\{\tau_1, \ldots, \tau_q\}$ so that the parameters to be estimated are $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2, \boldsymbol{\tau}^T)^T$, where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_q)^T$.

3   Parameter Estimation

A joint iterative process for estimating the fixed parameters and variance components is given by

$$\boldsymbol{\beta}^{(r+1)} = \left[\sum_{i=1}^n v_i^{(r)} \mathbf{X}_i^T \boldsymbol{\Sigma}_i^{-(r)} \mathbf{X}_i \right]^{-1} \left[\sum_{i=1}^n v_i^{(r)} \mathbf{X}_i^T \boldsymbol{\Sigma}_i^{-(r)} \mathbf{y}_i \right] \quad (5)$$

and

$$\boldsymbol{\gamma}^{(r+1)} = \text{argmax}_{\boldsymbol{\gamma}} \{l(\boldsymbol{\beta}^{(r+1)}, \boldsymbol{\gamma}^{(r)})\}, \quad (6)$$

for $r = 0, 1, 2, \ldots$, where $\boldsymbol{\gamma} = (\sigma^2, \boldsymbol{\tau}^T)^T$, $v_i = -2\frac{g'(u_i)}{g(u_i)}$ and $l(\boldsymbol{\beta}, \boldsymbol{\gamma})$ denotes the log-likelihood function. As in the normal case we can consider the posterior distribution of $\mathbf{b}_i$ given the observed data $\mathbf{y}_i$ to estimate the unit-specific parameters $\mathbf{b}_i$'s, which is also an elliptical distribution (see, for example, Fang, Kotz and Ng, 1990). Thus, by assuming that $\boldsymbol{\Sigma}_i$ is known, the empirical Bayes estimate is given by

$$\hat{\mathbf{b}}_i = \text{E}\left[\mathbf{b}_i | \mathbf{Y}_i = \mathbf{y}_i, \hat{\boldsymbol{\beta}}, \boldsymbol{\gamma} \right] = \mathbf{D} \mathbf{Z}_i^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}). \quad (7)$$

The variance-covariance matrix of $\hat{\mathbf{b}} = (\hat{\mathbf{b}}_1^T, \ldots, \hat{\mathbf{b}}_q^T)^T$ takes the form

$$\text{Var}(\hat{\mathbf{b}}) = \Delta \mathbf{Z}^T \mathbf{\Sigma}^{-1} \text{Var}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \mathbf{\Sigma}^{-1} \mathbf{Z} \Delta, \tag{8}$$

where $\Delta = \mathbf{D} \otimes \mathbf{I}_{m_i}$, $\mathbf{Z} = \text{diag}(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$, $\mathbf{\Sigma} = \text{diag}(\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_n)$, $\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T)^T$ and $\mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T)^T$. The calculation of $\text{Var}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ involves the quantities v_i's and becomes more complicated than in the normal case. For v_i fixed, we have $\text{Var}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{\Sigma}^* \mathbf{Q}^* \text{Var}(\mathbf{y}) \mathbf{Q}^* \mathbf{\Sigma}^*$ where $\mathbf{\Sigma}^* = \text{diag}(v_1 \mathbf{\Sigma}_1, \ldots, v_n \mathbf{\Sigma}_n)$, $\mathbf{Q}^* = \left[\mathbf{\Sigma}^{*-1} - \mathbf{\Sigma}^{*-1} \mathbf{X} \left(\mathbf{X}^T \mathbf{\Sigma}^{*-1} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{\Sigma}^{*-1} \right]$ and $\text{Var}(\mathbf{y}) = \alpha \mathbf{\Sigma}$ with $\alpha > 0$ being a constant that may be obtained from the derivative of the characteristic function (see, for example, Fang, Kotz and Ng, 1990). For the Student-t distribution with ν degrees of freedom, for instance, $\text{Var}(\mathbf{y}) = \frac{\nu}{\nu-2} \mathbf{\Sigma}$. In practice, $\mathbf{\Sigma}_i$ is not known, and it is usual to replace it by its maximum likelihood estimate, as well as v_i.

4   Assessing Variance Components

Since in the marginal model (3) the parameters $(\tau_1, \ldots, \tau_q)$ are not required to be positive we can perform, for instance, a likelihood ratio test to assess $H_0 : \boldsymbol{\tau} = \mathbf{0}$ against $H_2 : \boldsymbol{\tau} \neq \mathbf{0}$. However, because the main interest is in one-sided alternatives and due to the simplicity of score tests, we will apply the score-type test proposed by Silvapulle and Silvapulle (1995) to assess $H_0 : \boldsymbol{\tau} = \mathbf{0}$ against $H_1 : \boldsymbol{\tau} > \mathbf{0}$, with at least one strict inequality in H_1. This score-type test has been recently applied for assessing one-sided alternatives for dispersion parameters. For example, Paula and Artes (2000) use the score-type test to assess overdispersion in logistic regression models for grouped data, while Verbeke and Molenberghs (2003) discuss the application of the test in the assessment of variance components in normal mixed models. Consider the decomposition of the score function $\mathbf{S} = (\mathbf{S}_\lambda^T, \mathbf{S}_\tau^T)^T$ and the Fisher information matrix $\mathbf{K} = (\mathbf{K}_{\lambda\lambda}, \mathbf{K}_{\lambda\tau}, \mathbf{K}_{\tau\lambda}, \mathbf{K}_{\tau\tau})$ to conform with $\boldsymbol{\theta} = (\boldsymbol{\lambda}^T, \boldsymbol{\tau}^T)^T$, where $\boldsymbol{\lambda} = (\boldsymbol{\beta}^T, \sigma^2)^T$. The score-type test is given by

$$\text{T}_S = \tilde{\mathbf{Z}}^T \mathbf{K}_{\tau\tau}^{-1} \tilde{\mathbf{Z}} - \inf_{[\mathbf{a} \geq \mathbf{0}]} \{ (\tilde{\mathbf{Z}} - \mathbf{a})^T \mathbf{K}_{\tau\tau}^{-1} (\tilde{\mathbf{Z}} - \mathbf{a}) \}, \tag{9}$$

where $\tilde{\mathbf{Z}} = [\tilde{\mathbf{S}}_\tau - \tilde{\mathbf{K}}_{\tau\sigma^2} \tilde{\mathbf{K}}_{\sigma^2\sigma^2}^{-1} \tilde{\mathbf{S}}_{\sigma^2}]$, with all the quantities evaluated at the null estimate $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}^T, \tilde{\sigma}^2, \mathbf{0}^T)^T$. Under suitable regularity conditions and for large n, one has that $\text{T}_S \overset{\text{H}_0}{\sim} \sum_{\ell=0}^q \omega(\ell; \boldsymbol{\Delta}) \chi_\ell^2$, where χ_0^2 denotes the degenerate distribution at the origin, $\boldsymbol{\Delta} = \text{Var}(\hat{\boldsymbol{\tau}})$ and $\omega(\ell; \boldsymbol{\Delta})$'s are known as level probabilities and are expressed as functions of correlation coefficients associated with the $q \times q$ matrix $\boldsymbol{\Delta}$ (see, for instance, Shapiro, 1985).
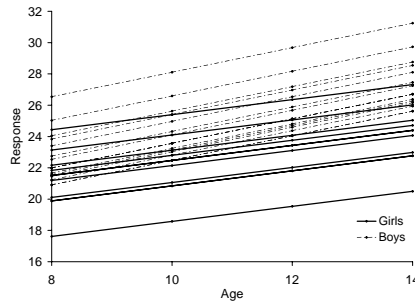
FIGURE 1. Individual adjusted profiles for the Student-t model.

5   Application

By way of illustration, we will consider the orthodontic data set presented by Potthoff and Roy (1964), where the response variable is the distance (in millimeters) between the pituitary and the pterygomaxillary fissure, which was measured at 8, 10, 12, and 14-years-olds in two groups, boys and girls. We fitted several models in order to apply the statistic T_S in different situations: first, by assuming a multivariate normal distribution, and second, by assuming a multivariate Student-t distribution with 6 degrees-of-freedom for the boys' group, as suggested by Pinheiro et al. (2001), and with 30 degrees-of-freedom for the girls' group. The independence model was tested against (i) the one with random intercept, (ii) the random slope model, and (iii) the model that includes these two random effects. For all these three situations, the null hypothesis was rejected. For the normal models, the values of T_S were, respectively, 61.8, 58.5 and 46.5 while for the Student-t models, they are equal to, respectively, 62.5, 61.0 and 50.9. One more situation was considered in which random slope effect was tested under the presence of random intercept effect, and the results for the T_S statistic were 0.60 under the normal distribution and 1.86 under the Student-t distribution. Therefore, the conclusion for the normal and Student-t models was that the final model should include only the random intercept. presents the parameter estimates and their approximate standard errors, which, under the t-model, are smaller than under the normal model. As pointed out by Pinheiro et al. (2001), two boys were identified as outliers under the normal model. The influence of dropping these observations on the parameter estimates was evaluated. Variations on the parameter estimates were in general smaller under the Student-t model, confirming the robustness of this model against extreme points, even though the inferential results remain unchanged. The influence of dropping the outlying observations on T_S was also evaluated, and the results showed that variations on T_S were also smaller under the Student-t model. describes the individual adjusted

TABLE 1. Parameter estimates for the random intercept models.

Group	Parameter	Normal Estimate (st.-error)	Student-t Estimate (st.-error)
Boys	Intercept	16.34 (0.96)	16.93 (0.84)
	Slope	0.78 (0.08)	0.72 (0.06)
Girls	Intercept	17.37 (1.16)	17.43 (0.95)
	Slope	0.48 (0.09)	0.47 (0.07)
	σ^2	1.87 (0.29)	1.04 (0.20)
	τ	3.03 (0.96)	2.85 (0.94)

profiles for the Student-t model with random intercept.

References

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

Laird, N. M. and Ware, J. H. (1982). Random effect models for longitudinal data. *Biometrics* **38**, 963-974.

Paula, G. A. and Artes, R. (2000). One-sided test to assess correlation in linear logistic models using estimating equations. *Biometrical Journal* **42**, 701-714.

Pinheiro, J.C., Liu, C. and Wu, Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effect models using the multivariate *t* distribution. *Journal of Computational and Graphical Statistics* **10**, 249-276.

Potthoff, R. F. and Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve models. *Biometrika* **51**, 665-680.

Shapiro, A. (1985). Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika* **72**, 133-144.

Silvapulle, M. J. and Silvapulle, P. (1995). A score test against one-sided alternatives. *Journal of American Statistical Association* **90**, 342-349.

Verbeke, G. and Molenberghs, G. (2003). The use of score test for inference on variance components. *Biometrics* **59**, 254-262.

Modelling Financial Durations between Price Movements

Giovanni De Luca[1] and Giampiero M. Gallo[2]

[1] Istituto di Statistica e Matematica, Università di Napoli "Parthenope", Via Medina 40 - 80133 Napoli, Italy - `giovanni.deluca@uninav.it`
[2] FEDRA and Dipartimento di Statistica "G.Parenti", Università di Firenze, Viale G.B. Morgagni, 59 - 50134 Firenze, Italy – `gallog@ds.unifi.it`

Abstract: In this paper we apply an autoregressive conditional duration model discussed in De Luca and Gallo (2004) to a long series of observations from the transactions on the IBM stock in April 2001. We show that the restriction imposed by a simple exponential distribution for the innovation term is too binding and that a mixture–based approach delivers a better fit and a wider array of interpretation of the results.

Keywords: Ultra–high frequency data; Autoregressive Conditional Duration Models; Market microstructure; Mixture of distributions.

1 Introduction

Movements of asset prices in financial markets are the focus of quantitative analysis in order to recognize patterns in their functioning. The goal is to study the behavior of markets, to analyze the features of exchanges, to provide explanations and possible guidelines to the evolutions in the future. Among the objects of analysis so called financial durations, i.e. the time distance between events of interest (a single trade, the accumulation of a certain amount of traded volume, the movement of an asset price above or below a certain threshold), have recently gained increasing attention among practitioners and academicians alike. This interest was made possible by the recording and diffusion of ultra-high frequency data (Dacorogna *et al.*, 2001), that is data that collects all transactions about a trade as it occurs (including the time at which this occurred, the volume exchanged and the price at which the asset was sold or bought) and the development of a new branch of econometrics, Engle (2000). Not all price movements are relevant: as a matter of fact since assets are usually quoted at two prices, a bid (i.e. the highest price somebody is willing to pay to buy the asset) and an ask price (i.e. the lowest price somebody is willing to receive to sell the asset), the observed series of traded prices reflect the fact that market makers are at times counterparts to a buy trade, at others to a sell trade. It becomes therefore of interest to analyze the duration between
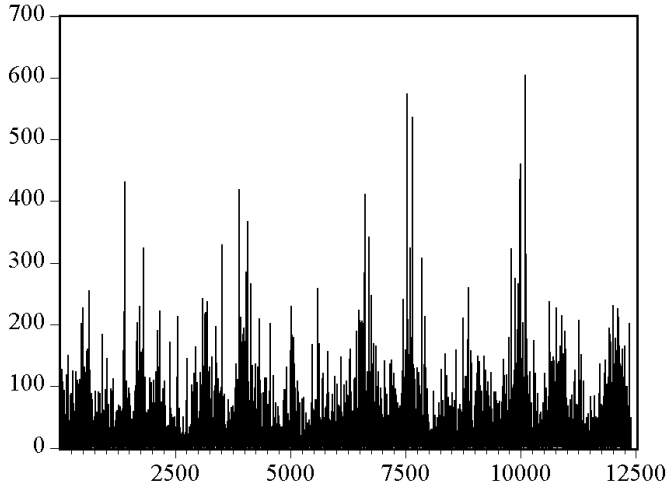
FIGURE 1. IBM stock. Durations between price movements above $ 0.0625.

meaningful movements of prices (either up– or downward) above a certain threshold. Furthermore since the observed series that ensues is irregularly spaced, new models are required to represent these data satisfactorily. The class of Autoregressive Conditional Durations (ACD) models put forth by Engle and Russell (1998) aims at reproducing the stylized facts of duration clustering the same way as the famed GARCH models (Bollerslev *et al.*, 1994) aim at modelling financial volatility clustering.

We will start by presenting two models both based on exponential innovations estimated on data related to a few days of transactions for the IBM stock (12399 observations selected in correspondence to price movements above 1/16th of a US dollar in module and after adjusting for errors in the data, cf. the pattern of the data in Figure 1). The first model is the ACD with exponential errors and the estimated results point out that some of the features of the model do not fit well the characteristics of the data, namely the variance of the estimation residuals is far from the theoretical one. The second model is a modification of the ACD and it is called Mixture–based ACD (MACD) discussed in detail in De Luca and Gallo (2004). The empirical results show that the MACD is capable of a better fit better, especially in capturing a higher variance in the data.

2   The Models

Let X_i be the duration between two movements in price beyond a certain threshold occurred at times t_{i-1} and t_i. Apart from some intra–daily sea-

sonal component (cf. Engle and Russell, 1998, among others, characterized by market microstructure problems, such as different speed of activity at opening, lunch and closing times or following some news release) which can be removed (for details cf. De Luca and Gallo, 2004) producing a "clean" duration x_i which can be modelled as a Multiplicative Error Model (MEM, Engle, 2002; Engle and Gallo, 2004):

$$x_i = \Psi_i \epsilon_i \tag{1}$$

$$\Psi_i = \omega + \sum_{j=1}^{q} \alpha_j x_{i-j} + \sum_{j=1}^{p} \beta_j \Psi_{i-j}, \tag{2}$$

$$\epsilon_i \sim \text{iid exponential}(1). \tag{3}$$

The specific model is called an $\text{ACD}(p,q)$ with exponential errors with suitable conditions on the parameters in order to ensure stationarity and a strictly positive conditional expected duration.

Rather than modifying the structure of the conditional expected duration Ψ_i as in other contributions in the literature (cf. the references in De Luca and Gallo, 2004), one can intervene on the nature of the innovation term. Beside the Weibull distribution, a promising process for ϵ_i is one in which there is a mixture of two exponential distributions with a weight $0 < p < 1$ attributed to one and the complement $1 - p$ to the other:

$$f\left(\epsilon_i; \mathcal{I}_{i-1}\right) = p f_1\left(\epsilon_i; \theta_1, \mathcal{I}_{i-1}\right) + (1-p) f_2\left(\epsilon_i; \theta_2, \mathcal{I}_{i-1}\right). \tag{4}$$

The parameters θ_1 and θ_2 characterize the pdf's of either distribution. While we still need the expectation of the mixture–based innovation term to be unit (and accordingly we impose appropriate constraints), the two exponentially distributed components have instantaneous rate of transaction different from one another. The important feature of this specification is that the variance of the innovation is greater than one, departing in a substantial manner from the simple exponential case. The weight p can be conveniently interpreted in reference to the price formation mechanism and the presence of different types of traders in the market. The term Ψ_i retains the interpretation of modelling the expected conditional duration in an autoregressive manner to capture persistence.

3   The Data and the Results

As mentioned before, for reasons of space we concentrated on a single blue chip stock, IBM: the chosen period spans from Apr. 1, 2001 to Apr.17, 2001. The transaction data was extracted from the Trades and Quotes database of the NYSE. After seasonally adjusting the data for time–of–the–day effects with a cubic spline with nodes set every hour, the 12399 observations were used to estimate the unknown parameters by QML. The

TABLE 1. QML Estimates of ACD models.

Parameter	ACD(1,1)	ACD(1,2)	MACD(1,1)	MACD(1,2)
ω	0.1620	0.1569	0.2486	0.2369
	(0.0231)	(0.0222)	(0.0400)	(0.0345)
α	0.1130	0.1306	0.1337	0.1501
	(0.0097)	(0.0101)	(0.0139)	(0.0135)
β_1	0.7261	0.3437	0.6187	0.3027
	(0.0308)	(0.0464)	(0.0501)	(0.0562)
β_2	-	0.3699	-	0.3085
	(-)	(0.0479)	(-)	(0.0560)
p_1			0.6458	0.6456
			(0.0160)	(0.0161)
λ_1			0.4729	0.4745
			(0.0135)	(0.0136)
Diagnostics				
$Q(15)$	36.572	25.825	37.7407	24.250
p-value	0.00146	0.0400	0.0010	0.0610
Mean	1.000	1.000	1.000	1.000
p-value	0.9911	0.9868	0.9879	0.990
Variance	2.074	2.060	2.101	2.081
p-value	0.000	0.0000	0.2450	0.3213
log-likelihood	-12032.48	-12009.38	-11190.18	-11177.47
Theoretical Var	1	1	2.013	2.006

results are presented in Table 1 for the simple exponential and the mixture–based exponential cases and for the specification (1,1) and (1,2). Below the parameter estimates we report the standard errors.

Some comments are in order: first of all the diagnostics on the autocorrelation of estimated residuals as shown by the Ljung Box statistic is still a problem. The second feature of the results is that the theoretical variance equals one in the standard case, whereas the estimated variance of the residuals is always above the value of 2. A better fit is had by the mixture based model where next to the significance of all parameters we notice the important result of the variance of the residuals being close to the theoretical value implied by the model (computed from the estimated parameter values). The log-likelihood values also signify a much better fit of our proposal relative to the base case.

4   Conclusions

In this paper we have shown the empirical superiority of a mixture–based approach to modelling financial durations between price changes above a certain threshold. Coupled with removal of intradaily systematic patterns

of trading, such a strategy allows one to concentrate on modelling the time elapsed between meaningful market movements. For reasons of space, many issues remain undiscussed such as the sensitivity of the modelling effort to the size of the threshold and to the type of seasonal adjustment procedure. As discussed in De Luca and Gallo (2004), the mixture–based approach needs to be extended in the direction of allowing the weights of the mixture to be variable, possibly as a function of variables in the information set.

References

Bollerslev, T., R.F. Engle and D.B. Nelson (1994) ARCH Models. Chapter 49 of *Handbook of Econometrics* Volume IV, (ed. by R.F.Engle and D. McFadden). Amsterdam: Elsevier Science, 2961-3031.

Dacorogna, M.M., Gençay, R., Müller, U., Olsen, R.B. , Pictet O.V. (2001). *An Introduction to High-Frequency Finance.* San Diego: Academic Press.

De Luca, G., and G.M. Gallo (2004). Mixture Processes for Intradaily Financial Durations. Forthcoming in *Studies in Nonlinear Dynamics and Econometrics.*

De Luca, G., and P. Zuccolotto (2003). Finite and infinite mixtures for financial durations. *Metron*, **61**, 431455.

Engle, R.F. (2000). The Econometrics of Ultra-High Frequency Data, *Econometrica*, **68**, 1-22.

Engle, R.F. (2002). New Frontiers for ARCH Models, *Journal of Applied Econometrics*, **17**, 426-446.

Engle, R.F., and G.M. Gallo (2003). A Multiple Indicators Model for Volatility Using Intra-Daily Data, *NBER Working Paper 10117.*

Engle, R.F., and J.E. Russell (1998). Autoregressive Conditional Duration: a New Model for Irregularly Spaced Transaction Data, *Econometrica*, **66**, 1127-1162.

Wavelet analysis of electrical signals obtained from experimental design

A. Di Bucchianico[1], H.P. Wynn[2] and T. Figarella[1]

[1] EURANDOM, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
[2] EURANDOM and London School of Economics

Abstract: Nowadays, electronic products tend to be economically outdated before their technical end-of-life has been reached. The ability to analyze and predict the (remaining) technical life of a product would make it possible either to re-use sub-assemblies in the manufacture process of new products, or to design products for which the technical and economical life match. This requires models to predict and monitor performance degradation profiles. In this paper we report on designed experiments to obtain such models. We show how wavelet analysis can be used to extract features from electrical signals. These features are analyzed using the Analysis of Variance in order to establish relations between these features and performance degradation.

Keywords: Signature analysis; Wavelet analysis; Peak extraction; Analysis of variance.

1   Introduction

The context of this project is the current trend to assemble complex products from modules supplied by other companies. Signature Analysis is a technique that allows to measure the parameters, which are significant for the lifespan of complex products like copier machines. By means of SA the prediction of the lifetime is not 'failure-driven' but 'performance-driven'. In other words, Signature Analysis is not based on the measurements of undesirable or irregular functionality, but it predicts the lifespan on basis of the actual technical performance of a complex compound product of the system.

In this paper we show the results of the experiment performed in the sub-module Main tray of the finisher module (Figarella, 2003). Specifically, we only consider the *stapler motor*, which is one of the three parts of the Main tray involved in the experiment. The stapler motor stitches three staples in each piece of paper.

During the experiment five electrical signals, corresponding to current consumption of the stapler motor, are measured as responses per run in the experiment. The classical multivariate analysis cannot be performed with

signals as response variables because a signal is a function of a continuous variable instead of a value. Therefore, the *maximum amplitude* of the first peak of the current signal, is taken as a feature or characteristic see Figure 1. Afterwards, Analysis of Variance is performed using this value. Extracting the features manually, i.e., without any mathematical method, is time consuming and not very accurate. Since the nature of the information contained in the signals is local we have chosen wavelet analysis over time series to obtain reliable features.

The experiment on the main tray is a replicated 2^{7-3} fractional screening experiment, where the seven factors obtained from a so-called *Failure Mode Effect Analysis* vary systematically. The objective of the experiment is to identify the influence of these factors on the features or characteristics extracted from the current signals. The results of this experiment will be input for further tests to obtain precise functional relationships. The final result is a monitoring scheme with limits for dominant parameters.

2   Wavelet Approach for analysis of stapler motor data

We have chosen wavelet analysis (Burrus, 1998 and Walnut, 2002) because it enables the analysis of localized areas of a larger signal. We assume that the behaviour of the replicated signals within a run is in general the same because they are generated by the same setting. By means of wavelet analysis we first simplify the description of a signal in terms of a small number of wavelet coefficients, and afterwards we use them as features to perform the Analysis of Variance.

We are interested in finding the maximum amplitude of the first peak of the current signal of the stapler motor (see Figure 1). This peak measures the current consumption during the action *spring load*; at this point the stapler anvil goes down against the paper.

We start the exploratory analysis with a pre-processing step in order to get rid of part of the noise, through spectral analysis of the signal, and then we apply the wavelet theory to obtain the features. Since the signal was oversampled we decided to downsample the signal, and in this way we reduce computational time by removing part of the high frequency component of the signal. We have carried out three wavelet-based approaches to obtain the features. The first one was the so-called *level-dependent thresholding* (Jansen, 2001). However, we do not present the results of this approach in this paper because it did not work properly because the noise seems to be non-gaussian. In the following we present the results of the other two approaches used after downsampling the signals.
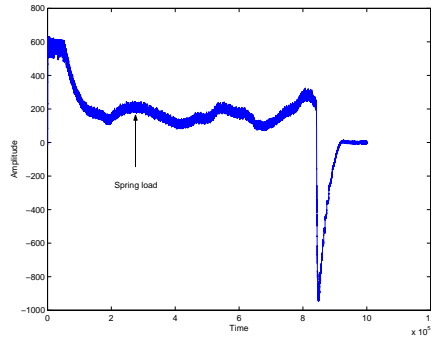
FIGURE 1. Current signal of the stapler motor and spring load peak

2.1   Approach 1: Rough denoising - Extracting the features using A_6

Rough denoising consists of decomposing the signal at several levels, removing all high-frequency components at each level, and then reconstruct the signal. Afterwards, we obtain a smooth signal and we extract the maximum of the first peak. At the 6th approximation level, A_6, almost no noise is present and it still keeps the main features of the signal visualizing the strength of the wavelet analysis, see Figure 2. At scales finer than level 6, there is little contribution to the signal. Therefore, the features are extracted from approximation level A_6.
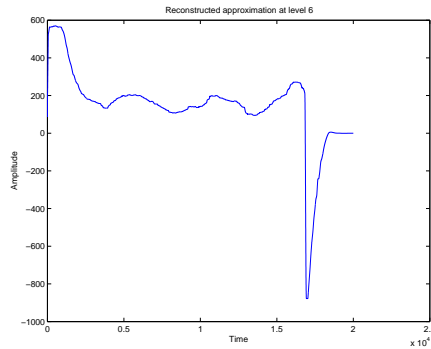
FIGURE 2. Reconstructed approximation at level 6

TABLE 1. Summary of the ANOVA results for the spring load peak

Factors and interactions	Manual extraction	Approach 1	Approach 2
Supply voltage 24 Vdc	0.00	0.00	0.00
Number of sheets	0.28	0.00	0.00
Feed roll load	0.40	0.82	0.79
PWBA modification	0.00	0.00	0.00
PWBA temperature	0.08	0.69	0.76
Belt tension	0.01	0.65	0.60
Supply voltage 5 Vdc	0.85	0.26	0.20
Supply 24 Vdc:number of sheets	0.08	0.02	0.00
Supply 24 Vdc:feed roll load	0.47	0.73	0.41
Supply 24 Vdc:PWBA modification	0.19	0.31	0.43
Supply 24 Vdc:PWBA temperature	0.41	0.38	0.38
Number of sheets:feed roll load	0.15	0.07	0.08
Number of sheets:PWBA modification	0.79	0.10	0.96
Feed roll load:PWBA modification	0.02	0.32	0.32
Residual standard error	9.86	6.10	5.69

2.2 Approach 2: Extracting the features using the average of approximation coefficients

In this approach we work directly on the wavelet coefficients without reconstruction. While we increase the level of decomposition, the length of the coefficient vector is halved. For example, the length of the approximation coefficients at level 4 is slightly more than $1/2^4$ the length of the downsampled signal. Therefore, at level 8 we have represented the complete signal by only few coefficients, approximately 95 coefficients.

After extracting the wavelet coefficients in each level, we calculate the maximum of the first peak of the coefficients at levels 4 up to 8. Then we calculate the weighted average of the maximum of the wavelet coefficients of each level. The weights are given by $2^{-j/2}$ for levels $j = 4, \ldots, 8$, so the maximum of the different levels are at the same scale.

2.3 Results

Table 1 is a summary of the ANOVA for the first peak using the features extracted manually, the first and the second wavelet approach. The table contains the factors and interactions with their respective *P-values* (for simplicity F values are omitted). We see that few factors affect the maximum amplitude of the first peak. This is favourable for translating this peak back to internal degradation parameters of the machine, which is subject of future research. Taking the average of the maximum of the wavelet coefficients of the 5 levels we obtain the same significant factors and interactions as with the first approach. Furthermore the residual standard error is 42 times smaller than the residual standard error obtained with the manually extracted features.

3 Conclusions

We used several wavelet-based approaches but only two of them gave satis-factory features from the signals. The first approach is based on the recon-structed approximation at level 6 because it contains much less noise than the original signal, and it still keeps the main characteristics of the signal. In the second approach we use directly the wavelet coefficients at 5 levels and we average them.

For the first peak of the stapler motor, averaging the maxima of the wavelet coefficients appears to be the best approach since the residual standard er-ror is the smallest, and because it considers the information from several levels of decomposition assuring stability of the feature. Besides the re-duction of the residual standard deviation and the number of outliers, the computation time during the wavelet analysis is negligible. Therefore, our method can be used for on-line extraction of signal features.

References

Burrus, C.S., Gopinath, R.A., Guo, H. (1998). *Introduction to Wavelets, and Wavelet Transforms: A Primer.* Upper Saddle River: Prentice Hall.

Figarella, T. (2003). ASREML User's Manual. New South Wales Agricul-ture.

Jansen, M.H. (2001). *Noise Reduction by Wavelet Thresholding.* New York: Springer.

Walnut, D.F. (2002). *An Introduction to Wavelet Analysis.* Boston: Birk-häuser.

The Shifted Warped Normal Model for Mortality

Paul H. C. Eilers[1]

[1] Department of Medical Statistics, Leiden University Medical Centre

Abstract: Age distributions of deaths due to specific diseases show strong skewness to the left. Using P-splines, an transformation of age is computed in such a way that the distributions become normal, but shifted over time on the transformed scale. The model is illustrated with data on deaths from respiratory diseases in the USA.

Keywords: Functional data analysis; Life table; P-splines.

1 Introduction

Human mortality shows complicated and interesting patterns. More and more data become easily available through the Internet, offering fascinating possibilities for data analysis and statistical modelling. Here I report on experiments with mortality data — more precisely, counts of deaths — from the United States. In each year from 1959 to 1998, the number of people dying from respiratory diseases are given in one-year intervals, separately for men and women.
The frequency distributions are skew with a long left tail. The right tail tends to become shorter over the years and the position of the peak shifts to the right. A normal distribution is certainly out of the question. But can we find a transformation of the age axis such that the distribution becomes essentially normal? The answer will be shown to be affirmative. On the transformed ("warped") scale the changes from year to year correspond to a shift, a change in the mean of the distribution. The optimal transform of the age axis is estimated with P-splines.

2 The shifted warped normal model

Figure 1 gives an impression of the number of deaths due to respiratory diseases, separately for men and women. Totals per year, summed over ages from 21 to 120 are presented, as well as age distributions for selected years.. The overall level has increased strongly over the years, especially for women. The age at which the peak occurs has shifted to the right, especially for men, while the right tail has become shorter.
It would be attractive to have a transformed age scale, such that the distributions would have the shape of the normal distribution. The changes between the years would then correspond to shifts in their means. If we let
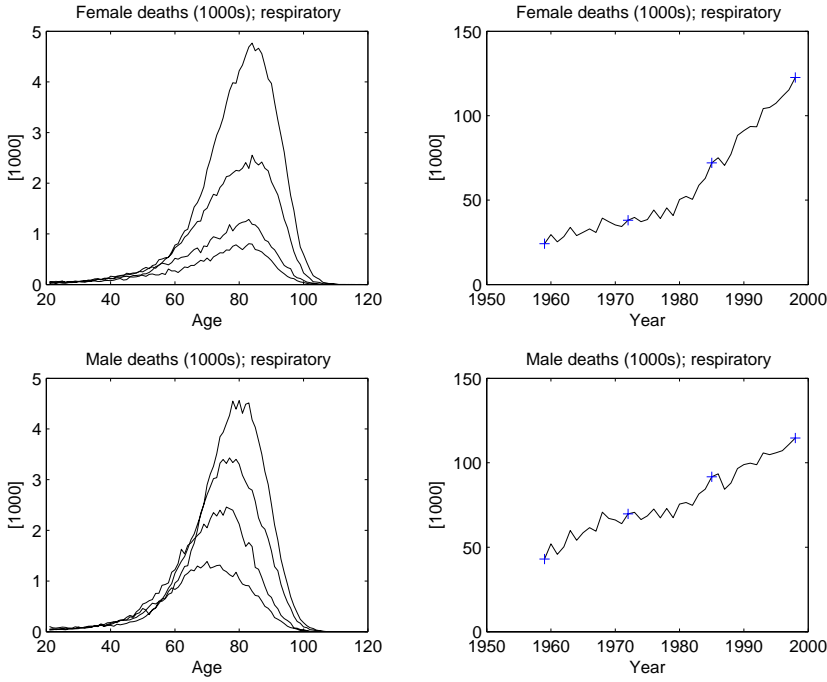
FIGURE 1. Deaths in the USA due to respiratory diseases, for women (top) and men (bottom). The crosses in the right panels indicate for which years the distributions are presented in the left panels.

y_{ij} indicate the value of the scaled distribution (i.e. divided by its maximum) at age a_i in year t_j, then the proposed model is:

$$\mu_{ij} = E(y_{ij}) = f(g(a_i) - \beta_j), \quad \text{with} \quad f(u) = \exp(-u^2/2). \qquad (1)$$

The unknown curve $g(a)$ is modelled in a semi-parametric way as a sum of B-splines in a:

$$g(a) = \sum_{k=1}^{K} B_k(a)\alpha_k. \qquad (2)$$

In the spirit of P-splines, the number of basis function, K, is relatively high (about 10) and a roughness constraint on the coefficient vector α is used to tune smoothness (Eilers and Marx, 1996). The following penalized sum of squares goal is minimized:

$$S = \sum_i \sum_j (y_{ij} - \mu_{ij})^2 + \lambda \sum_k (\Delta^2 \alpha_k)^2. \qquad (3)$$
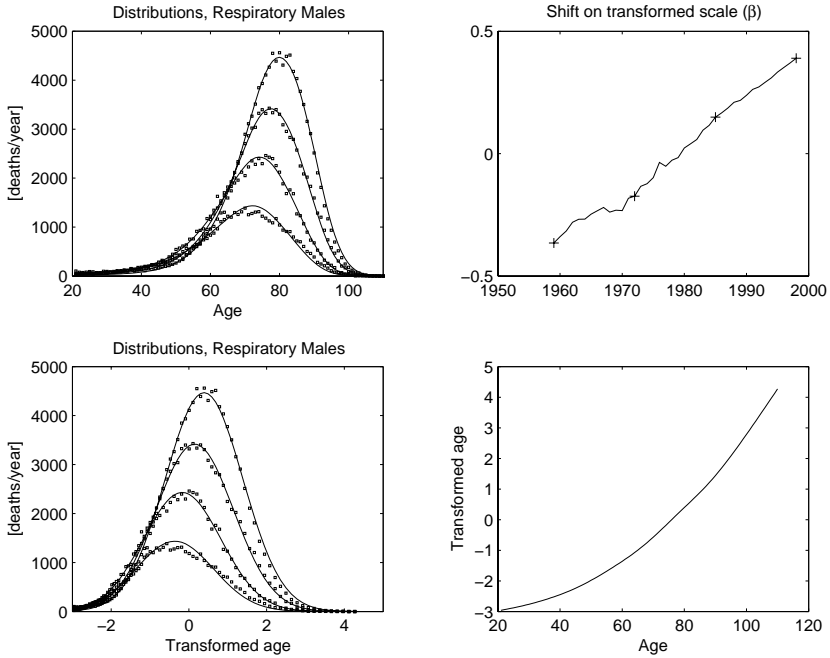
FIGURE 2. The fit of the model (lines) to the data (dots) for men. The left panels show distributions for the years that correspond to the crosses in the uper right panel.

It is clear that μ depends on α in a highly non-linear way, because they appear in the argument of the function f. Using a first-order Taylor expansion it can be linearized and proper starting values are easily found. To start the transformation estimate, $\tilde{g}(a) = (a - 75)/15$ was used, and the starting value for β_j was minus weighted (by the age distribution) average of $\tilde{g}$ for year j. The value of λ did not have much influence on the estimated transform; $\lambda = 0.1$ was used to get the results presented here. The algorithm was implemented in Matlab and found to be stable and fast. Fitting takes a few seconds on an average PC.

Figure 2 shows the results of fitting the model to the data for men. Apparently a good fit is obtained and the estimated transformation shows strong curvature, rising steeper with increasing age. On the other hand, the graph of $\hat{\beta}$ vs. time is almost linear. Note that this is not forced by the model, it is a property of the data.

Figure 3 shows results for women. As indicated by the small trend in the shifts (β), they have shown little progress compared to men. The peak of their age distributions have hardly shifted over the years.
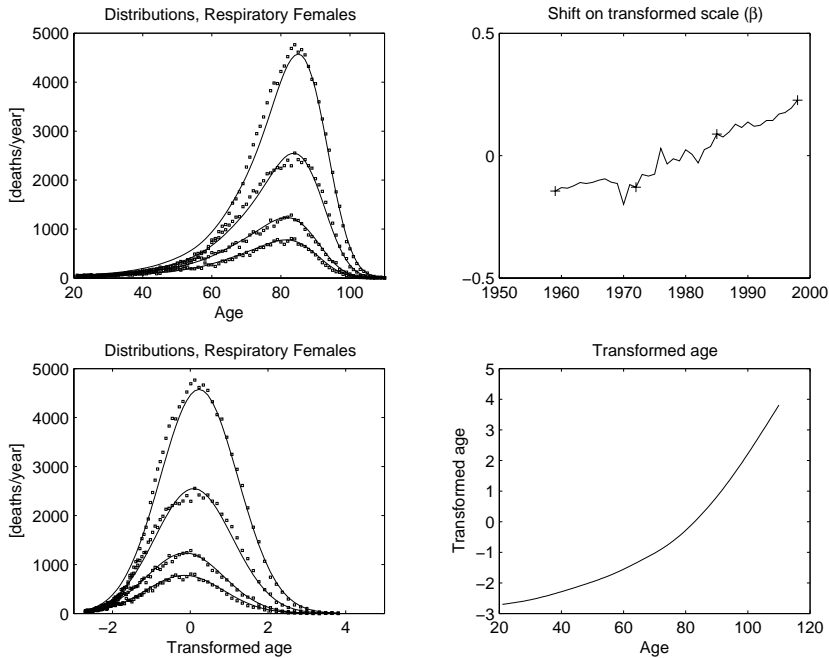
FIGURE 3. The fit of the model (lines) to the data (dots) for women. The left panels show distributions for the years that correspond to the crosses in the upper right panel.

3   Discussion

The shifted warped normal (SWaN) model appears to work well and easy to estimate. Still, on the technical front there is a lot to be improved. The least squares criterion, applied to the scaled distributions, is not optimal. By introducing an offset for each year, the model can be reformulated as a penalized GLM with a Poisson response and an unusual link function: the normal curve. The scoring algorithm can be applied for fitting it. A program for this algorithm has just been finished and seems to work well. Another question to be addressed is density correction with $|g(a)|$, because we transform the variable, age a on which the age distribution of deaths is computed. Intervals of equal width on the age scale generally have different widths on the g scale. This has been neglected in the present model.

It will be interesting to apply the model to more diseases and to more countries, or to different states within a country, to see which patterns are stable and which vary.

The data are also available as monthly counts and so seasonal effects can be studied. Experiments indicate that there is a strong seasonal pattern in β.

There is also a seasonal pattern in the height of the distributions over age. This way, only two parameters for each month (one for height, the other for shift) give a quite precise summary of seasonal changes in a distribution over 90 age classes.

The almost linear pattern in $\hat{\beta}$ suggests that the model lends itself well to extrapolation. This needs further research, e.g. using part of the data, say up to 1990, to "predict" the years that follow and check this with cross-validation.

This model has strong similarities to Functional Data Analysis (Ramsay and Silverman, 1997). They also align curves by scaling of the independent axis. But here an additive model for age and time is used in the argument of a pre-specified function (the normal curve).

Preliminary experiments have shown that the model also works for overall mortality, even over long periods (a century or more), if the age range is limited to 70 and over. Experiments are going on to compare different countries. The website www.mortality.org is a very rich source of high-quality data.

A remarkable outcome is that the standard deviations of the distributions, over transformed age, are constant over time. Of course, this is specified by the model, but there are no indications that a richer model is needed for a good fit to the data.

References

Eilers, P.H.C. and Marx, B.D. (1996) Flexible smoothing using B-splines and penalized likelihood (with Comments and Rejoinder). *Statistical Science*, **11** 89–121.

Ramsay, J.O., Silverman, B.W. (1997) *Functional Data Analysis.* Springer.

Structured additive regression for multicategorical space-time data: A mixed model approach

Thomas Kneib[1] and Ludwig Fahrmeir[1]

[1] Department of Statistics, University of Munich

Abstract: In many practical situations, simple regression models suffer from the fact that the dependence of responses on covariates can not be sufficiently described by a purely parametric predictor. For example effects of continuous covariates may be nonlinear or complex interactions between covariates may be present. A specific problem of space-time data is that observations are in general spatially and/or temporally correlated. We propose a general class of structured additive regression models (STAR) for multicategorical responses, allowing for a flexible semiparametric predictor. This class includes models for multinomial responses with unordered categories as well as models for ordinal responses. We present our approach from a Bayesian perspective, allowing to treat all functions and effects within a unified general framework by assigning appropriate priors with different forms and degrees of smoothness. Inference is performed on the basis of a multicategorical linear mixed model representation. Variance components, corresponding to inverse smoothing parameters, are then estimated by using restricted maximum likelihood.

Keywords: Multicategorical space-time data; generalized linear mixed models; semiparametric regression; P-splines; restricted maximum likelihood.

1 Structured additive regression

Space-time regression data usually consist of a number of repeated observations on a response variable and a set of covariates, e.g. continuous covariates, categorical covariates, time scales, location indices or cluster indices. Different types of models have been introduced to analyze such data, depending on the type of the covariates and the distribution of the response. In many situations a purely parametric regression model is unable to describe the dependence of responses on covariates sufficiently. For example effects of continuous covariates may be non-linear or complex interactions between covariates might be present. A specific problem of space-time data is that observations may be spatially and/or temporally correlated. Within a parametric modelling framework, it is virtually impossible to include these aspects.

In recent years, models for space-time data and univariate responses have gained considerable attention (e.g. Kammann and Wand, 2003 or Fahrmeir, Kneib and Lang, 2004). However, the literature dealing with models for multicategorical space-time data is rather limited (compare Fahrmeir and Lang, 2001, for a notable exception based on Markov Chain Monte Carlo techniques and latent utilities). We propose a general class of structured additive regression models (STAR) for multicategorical responses, allowing for a flexible semiparametric predictor. This class includes models for multinomial responses with unordered categories as well as models for ordinal responses.

For ordinal responses we assume a cumulative regression model, i.e. the probability for observation y_{it}, $i = 1, \ldots, n$, $t = 1, \ldots, T$ to be in category r or less is assumed to be

$$P(y_{it} \leq r) = F(\theta_r - \eta_{it}), \qquad (1)$$

where F denotes a cumulative distribution function, e.g. the logistic or the standard normal distribution function, and $\theta_1 < \ldots < \theta_q$ are ordered thresholds. Nominal responses can be analyzed using multinomial logit models but we will focus on the ordinal case here (compare Kneib and Fahrmeir, 2004, for a more detailed description of both cases). For a space-time main effects model the semiparametric predictor η_{it} in (1) can be defined by

$$\eta_{it} = f_1(x_{it1}) + \ldots + f_l(x_{itl}) + f_{time}(t) + f_{spat}(s_i) + u'_{it}\gamma, \qquad (2)$$

where, f_{time} and f_{spat} represent possibly nonlinear effects of time and space, $f_1, \ldots, f_l$ are unknown smooth functions of the continuous covariates $x_1, \ldots, x_l$, and $u'\gamma$ corresponds to the usual parametric linear part of the predictor. This model can be extended in various ways, e.g. to include interactions or individual-specific effects, compare Kneib and Fahrmeir (2004) and the example below. Note, that the observations y_{it} are marginally correlated, especially over time and space, but are assumed to be independent conditional on the effects in (2).

As an example, we analyze data from a forest health survey, where for several years the damage state of a population of trees is measured in three ordered categories. In addition to the continuous covariate age of the tree A and a vector of further (mostly categorical) covariates u, the location s of each tree is available on a lattice map. Due to the space-time structure of the data, we have to take temporal as well as spatial correlations into account. This can be achieved using a semiparametric predictor of the form

$$\eta_{it} = f_A(A_{it}) + f_{time}(t) + f_{time,A}(t, A_{it}) + f_{spat}(s_i) + u'_{it}\gamma. \qquad (3)$$

Here, the model in (2) is extended to include an interaction surface $f_{time,A}$ between calendar time and the age of the tree. Figure 1 shows estimates for the functions f_A, f_{time} and f_{spat}.
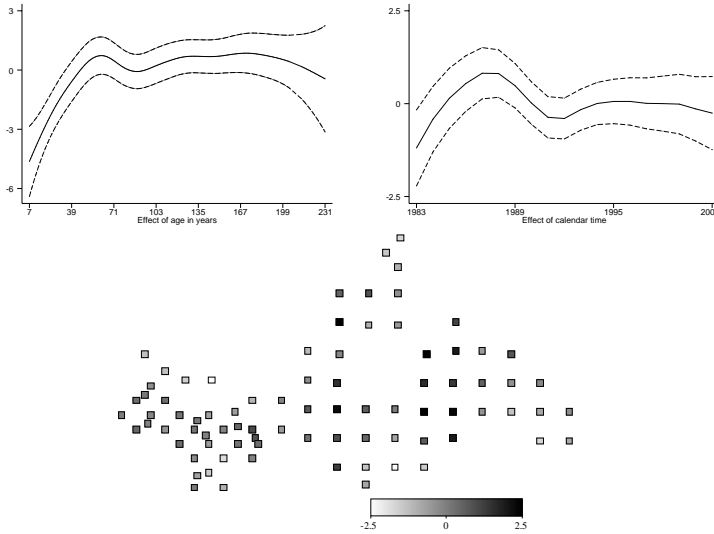
FIGURE 1. Estimated main effects of the age of the tree and calendar time together with pointwise 95% credible intervals and estimated spatial effect.

2   Prior assumptions

The Bayesian model formulation is completed by specifying appropriate priors for the different effects or, more specifically, for the corresponding vectors of function evaluations f. In our approach we are always able to express these vectors as the product of a design matrix X and a vector of regression parameters β, i.e. we have

$$f = X\beta. \tag{4}$$

Now we can formulate a prior for f based on a prior for the vector of regression coefficients β. It turns out, that this prior also has a general form, which is given by

$$p(\beta|\tau^2) \propto \exp\left(-\frac{1}{2\tau^2}\beta' K\beta\right), \tag{5}$$

where K is a penalty matrix. The penalty matrix K and the design matrix X determine the general characteristics of the function, e.g. whether the function is continuous or whether it is differentiable. The variance parameter τ^2 corresponds to the inverse smoothing parameter in a frequentist approach and controls the trade-off between flexibility and smoothness. Let us now briefly describe some possibilities to model the effects in (2) and (3):

- $f_j(x_j)$ functions of continuous covariates: P-splines (Eilers and Marx, 1996, Lang and Brezger, 2004).

  - Approximate f_j by a B-spline with a large number of knots.
  - Define a random walk prior for the B-spline coefficients β.
  - The design matrix X contains evaluations of the basis functions at the observed values of x_j.
  - The penalty matrix is given by $K = D'D$ with first or second order difference matrix D.

- $f_j(x_{j_1}, x_{j_2})$ interaction surface: Two-dimensional P-splines (Lang and Brezger, 2004).

  - Use tensor products of one-dimensional B-splines as basis functions.
  - Define a two-dimensional random walk prior for β.

- $f_{spat}(s)$ spatial function of exact locations s: Stationary Gaussian random fields (Kammann and Wand, 2003, Kneib and Fahrmeir, 2004).

  - GRFs are surface smoothers based on special basis functions.
  - The penalty matrix K is defined by the correlation function of the GRF.

- $f_{spat}(s)$ spatial function of connected geographical regions s: Markov random fields.

  - Define appropriate neighborhoods for the regions s.
  - Assume that the expected value of $f_{spat}(s)$ is the (weighted) average of the function evaluations of adjacent regions.
  - The penalty matrix K has the form of an adjacency matrix.

3 Mixed model inference

Inference for STAR models can be performed on the basis of a multicategorical linear mixed model representation. Model components described by (4) and (5) can always be reexpressed in terms of a parameter vector with flat prior and a second parameter vector with i.i.d. Gaussian prior. This allows to rewrite STAR models as variance components models. The variance components, corresponding to inverse smoothing parameters, can then be estimated using mixed model methodology, especially restricted maximum likelihood, also termed marginal likelihood in the literature. Given estimates of the variance parameters, regression coefficients are estimated by a modified Fisher-scoring procedure. Since variance components are treated as unknown constants, our approach can be viewed as empirical

Bayes/posterior mode estimation and is closely related to penalized likelihood estimation in a frequentist setting. Numerically efficient algorithms, developed in Fahrmeir, Kneib and Lang (2004), allow the computation of the estimates even for fairly large data sets.

4   Conclusions

The presented approach has several advantages:

- It allows to deal with a very broad class of regression models, that even extends the presented models (2) and (3). For example we can directly incorporate random effects, varying coefficient terms and flexible seasonal components in our model.

- All model components are treated in a unified way conceptually, allowing compact presentation and easier implementation.

- Real data applications and simulation studies have provided evidence that the approach works considerably well in many situations compared with the fully Bayesian procedure of Fahrmeir and Lang (2001).

- Software for fitting the presented models is available in the public domain software package *BayesX*. Therefore the methodology can easily be used in other areas of research, e.g. the analysis of unemployment durations in microeconomics or in consumer choice models.

References

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

Fahrmeir, L., Kneib, T. and Lang, S. (2004). Penalized Structured Additive Regression for Space-Time Data: a Bayesian Perspective. *Statistica Sinica*, to appear.

Fahrmeir, L. , Lang, S.] (2001). Bayesian Semiparametric Regression Analysis of Multicategorical Time-Space Data. *Annals of the Institute of Statistical Mathematics*, **53**, 10–30

Kammann, E.E. and Wand, M.P. (2003). Geoadditive Models. *Journal of the Royal Statistical Society C*, **52**, 1–18.

Kneib, T. and Fahrmeir, L. (2004). Structured additive regression for multicategorical space-time data: A mixed model approach. SFB 386 Discussion paper 377, Department of Statistics, University of Munich.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, to appear.

Analyzing Plaid Designs using Mixed Models

Fotios Siannis[1] , Vernon T. Farewell[1]

[1] MRC Biostatistics Unit, Institute of Public Health, University Forvie Site, Robinson Way, Cambridge CB2 2SR, UK.

Abstract: In this paper we propose a way of analyzing data from a plaid square design using multilevel mixed models. In the case of normal outcomes, this can be seen as a generalization to ANOVA analysis, where covariates can be included and extensions to unbalanced designs can be considered. Furthermore, based on the analysis on mixed models, the analysis of non–normal data can be considered, although fitting these models using existing software might prove a real challenge.

1   Introduction

Plaid designs are not very common, but they seem very convenient in some contexts. They were briefly considered by Yates(1937) for field experiments, where additionally to the usual latin square structure, entire rows and/or columns were subject to the same treatment. They also appear in some medical experiments. Hence, there is a need for understanding and exploring the possible ways of analyzing data arising from such designs. They appear to be very useful when the nature of the experiment makes it reasonable to have treatments arranged in a systematic way. Cochran & Cox(1957) (§7.32) discuss strip-plot or criss-cross designs, which are special cases of the plaid design. They point out that although plaid designs sacrifice precision in the main effects, this is compensated by a higher precision in the interactions. Therefore, if interactions are of central interest, these designs appear to be more accurate than either randomized blocks or simple split-plot designs.

2   Model for normally distributed responses

2.1   The FACS Data

In this work we consider data from the experiment reported in Solomon *et.al.*(1997), where the Facial Action Coding System (FACS) was used as means of identifying the expression of pain by facial movement. A training program based on it was developed to train physicians to evaluate the amount of pain experienced by patients. These data relate to 74 occupational and physical therapy students (the raters) who were randomly

<pre>
                                    Subjects
                                    Factor B
                        Level 1   Level 2   Level b
                         - - -   - - -   - - -
                           m   m   ...   m
    Practitioners
                           |
            Level 1 | n
                           |

                           |
    Factor A   Level 2 | n
                           |

                           .
                           .
                           .
                           |
            Level a | n
                           |
</pre>

r replicates of this layout

TABLE 1. Replicated Plaid Design

assigned to training and no training groups (37 raters in each group). For each rater the data include their ratings on a descriptor scale of pain, of the pain experienced by eight patients who were observed on videotape as they underwent a standardized procedure to assess motion of a painful shoulder joint. Both active motions, performed by the patient without assistance, and passive motions, in which a therapist guided the patient's limb through its range of motion, were observed for each patient. These patients were a selected group from a previous study based on FACS, and consisted of four expressive and four unexpressive patients.

2.2   Design

These data were discussed by Farewell and Herzberg(2003), where an analysis of variance for plaid designs was given and where outcomes of interest were assumed to be normally distributed. The structure of a standard replicated plaid design, without a split plot component, is given in Table 1. In the physician-patients data, columns are associated with patients, divided into two levels (expressive and unexpressive), and the rows are associated with medical practitioners, also divided into two levels (trained

and untrained). More specifically, following the notation of Table 1, we have a=b=2, $n = 37$, $m = 4$ and since we have only one replication, $r = 1$. A mixed model for this layout can be written as

$$y_{ijkl(i)m(j)} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}$$
$$+ \epsilon_{l(i)k} + \epsilon'_{m(j)k} + \epsilon''_{l(i)j} + \epsilon'''_{m(j)i} + \epsilon''''_{l(i)m(j)k}, \tag{1}$$

where i indexes the levels of Factor A, j indexes the levels of Factor B, k indexes replicates, $l(i)$ indexes Raters within A and $m(j)$ indexes Subjects within B. The five error terms in the model are all normally distributed with mean zero and respective variances $\sigma^2_{P:AR}, \sigma^2_{S:BR}, \sigma^2_{PB:AR}, \sigma^2_{SA:BR}$, and $\sigma^2_{PS:ABR}$. This is a multilevel model, where at the lowest level (level 1) we have the observation, while at level 2 we have a cross–classification between raters and patients, which are nested within the training and the expressive groups respectively. Rasbash & Goldstein(1994) discuss mixed hierarchical models with cross–classified random structures. They demonstrate that a two–level additive variance component model with crossing at level 2 can be expressed as a model with a single level 2 unit nested within a single level 3 unit. This model can then be considered as a model with two levels, where the covariance matrix of the random terms takes a block diagonal form.

2.3 Results

We fit the model

$$y_{ijl(i)m(j)v} = \mu + \alpha_i + \beta_j + \rho_v + (\alpha\beta)_{ij} + (\alpha\rho)_{iv} + (\beta\rho)_{jv} + (\alpha\beta\rho)_{ijv}$$
$$+ \epsilon_{l(i)} + \epsilon'_{m(j)} + \epsilon''_{l(i)m(j)}, \tag{2}$$

which is slightly different than (1), using *PROC MIXED* in SAS, which handles mixed models with normally distributed outcomes. Since $k = 1$ there is no replication effect, while the split plot design of the data is represented by an additional fixed effect ρ, where v indexes the two possible outcomes (active and passive). In our attempt to reproduce the analysis presented by Farewell and Herzberg(2003) as closely as we can, we omitted two of the random terms that were included in model (1). These terms are pooled with $\epsilon''_{l(i)m(j)}$, giving a single random term with 510 df. In this way, *PROC MIXED* produces F–tests for the fixed effects using the appropriate error terms, as seen in Table 2. For example, to test the effect of expressiveness (second line in Table 2) the correct error term would be the one produced by the 'patients within expressiveness' random term $\epsilon'_{m(j)}$ with 6 df. In Table 2, DF1 presents the df on the numerator and DF2 the df on the denominator of the F–test. Additionally, the residual variance is estimated to be $\sigma^2_e = 5.66$. The three way interaction, which is of interest, is tested against the residual error term with 588 df and appears to be significant.

Effect	DF1	DF2	F Value	$\Pr > F$
GROUP	1	72	11.38	0.0012
EXPRESS	1	6	7.36	0.0350
GROUP*EXPRESS	1	510	5.13	0.0239
MOVEMENT	1	588	1336.03	$< .0001$
GROUP*MOVEMENT	1	588	7.71	0.0057
EXPRESS*MOVEMENT	1	588	764.35	$< .0001$
GROUP*EXPRESS*MOVEMENT	1	588	6.37	0.0118

GROUP=Training group for raters

TABLE 2. SAS output for the tests of the fixed effects

This way of modelling, not only reproduces the results obtained by Farewell and Herzberg (2003), but it can also be seen to allow extensions. Based on the mixed model, unbalanced explanatory variables and/or covariate structures can be easily incorporated. This is particularly useful in medical examples, where explanatory variables on the patient level are of varied types. We fitted an extended version of the FACS data, where one rater in the trained group and two patients in the unexpressive group where added to create an unbalanced data set.

3   Ordinal response model

In section 2, we considered the analysis of this mixed model with normally distributed responses using standard software for hierarchical models. The implementation of this model facilitates its generalization to generalized linear mixed models. This means, for example, that satisfactory analysis of ordinal response data, given appropriate choice of distribution for the error terms, could be considered. Brown & Prescott(1999) discuss mixed models for categorical data (ch. 4), pointing out the limitations in fitting these models using existing software.

The form of the model will be exactly as in (1), where the random terms can still be considered to be normally distributed. Currently, there are some widely used software that can handle non–normally distributed responses (commands *PROC NLMIXED* in SAS and *nlme* in S-Plus/R as well as MLWin, the GLLAMM package in Stata and MIXOR). The complicated structure imposed by the plaid design is not easily accommodated, however, since most of the above commands and packages do not allow for multilevel structures.

4   Conclusions

Plaid designs arise naturally in certain contexts, and hence there is a need to explore more about what they can offer to researchers. The absence of a

standard methodology for analyzing non–normal response data, the complicated structure of the designs and the lack of comprehensive (and easy to use) software to deal with them, are possibly some of the reasons why these designs are not more widely used. We propose the use of generalized linear mixed models to analyze data from plaid square designs. We have focused on the software available to fit such models, and discuss the limitations. Therefore, this work can be seen as a first step to solving some of the above problems.

References

Brown H. and Prescott R. (1999) *Applied Mixed Models in Medicine.* Statistics in Practice, Wiley.

Cochran W. and Cox G. (1957). *Experimental Designs, 2nd edition.* New York, Wiley.

Farewell V. T. and Herzberg A. M. (2003). Plaid designs for the evaluation of training for medical practitioners. *Journal of Applied Statistics*, **30(9)**, 957–965.

Rasbash J. and Goldstein H. (1994). Efficient Analysis of Mixed Hierarchical and Cross–Classified Random Structures Using a Multilevel Model. *Journal of Educational and Behavioral Statistics*, **19(4)**, 337–350.

Solomon P., Prkachin K., and Farewell V. (1997). Enhancing sensitivity to facial expression of pain. *Pain*, **71**, 279–284.

Yates F. (1937). Design and analysis of factorial experiments. *Technical Communication 35, Harpenden, UK, Commonwealth Bureau of Solis.*

Model Building and Interpretation of Ordinal Multilevel Random Effects Models with Exogeneity and Endogeneity

Antony Fielding[1], Neil Spencer[2]

[1] Deparment of Economics University of Birmingham United Kingdom
[2] Business School University of Hertfordshire United Kingdom

Abstract: We focus on multilevel random effects models for ordered response such as occur in educational achievement research. In model development changes in parameter values are difficult to compare because of implicit rescaling of parameters in the linear predictor. We combine a heuristic method to handle this with proposals for instrumental variable estimation when regressors are endogenous. Simulated and real educational data are used to evaluate these proposals.

1 Introduction

We model here ordered responses with multilevel random effects of the type $F^1\left(\gamma_{ij}^{(s)}\right) = \theta_s - \{(X\beta)_{ij} + u_{0j}\}$ such as appear in educational progress (Fielding (1999)). Here $\gamma_{ij}^{(s)}$ is the cumulative probability that student i in school j obtains grade s. F^1 will be the probit link though other forms may be noted. X is a matrix of regressors, u_{0j} is the random effect of the school and the θ_s (s = 1, 2, ..., k-1 where there are k categories) are thought of as cut-points of an underlying latent variable scale (with $\theta_1 < \theta_2 < \ldots < \theta_{k-1}$). We use macros for PQL2 estimation in MLwiN discussed by Fielding (2002). We desire to build models by extending the introduction of effects starting from a null model with no regressors. Each development of the model rescales parameters so that the latent variable level 1 variance is fixed at unity for the probit. This makes a comparison of all parameters in different extensions difficult (Snijders and Bosker, 1999). To facilitate these comparisons Fielding (2003) has used Conditional Mean Scoring (CMS) of categories for the null model to approximately identify scaling factors which can be applied to results. Here we additionally consider a situation with endogenous regressors may be related to the random part of the model, as might happen in educational settings (Spencer & Fielding, 2002). Instrumental variable (IV) methods to deal with the inconsistency of standard estimation in such situations have been fairly successful (Spencer and Fielding, 2002). The basics of IV estimation are well

TABLE 1. Mean and standard errors of parameter estimates from fifty simulated datasets

Coefficient	Values used in simulations	Method 1 CMS not used IV not used	Method 2 CMS used IV not used	Method 3 CMS not used IV used	Method 4 CMS used IV used
Cut-point 1	-2.150	-2.477(0.224)	-2.142(0.177)	-1.095(0.120)	-2.124(0.174)
Cut-point 2	-1.672	-1.920(0.222)	-1.660(0.175)	-0.858(0.112)	-1.663(0.169)
Cut-point 3	-1.194	-1.376(0.216)	-1.189(0.175)	-0.620(0.107)	-1.201(0.173)
Cut-point 4	-0.717	-0.831(0.207)	-0.717(0.173)	-0.380(0.105)	-0.733(0.184)
Cut-point 5	-0.239	-0.284(0.208)	-0.245(0.179)	-0.135(0.103)	-0.258(0.194)
Cut-point 6	0.239	0.268(0.201)	0.233(0.177)	0.113(0.098)	0.224(0.192)
Cut-point 7	0.717	0.824(0.197)	0.714(0.176)	0.363(0.091)	0.706(0.181)
Cut-point 8	1.194	1.378(0.202)	1.376(1.264)	0.611(0.095)	1.188(0.183)
Cut-point 9	1.672	1.949(0.199)	2.050(2.595)	0.863(0.098)	1.679(0.179)
Cut-point 10	2.150	2.503(0.190)	2.167(0.171)	1.104(0.099)	2.146(0.174)
Centred prior test	0.800	1.409(0.056)	1.219(0.043)	0.409(0.036)	0.795(0.053)
School variance	1.000	0.796(0.190)	0.598(0.149)	0.348(0.071)	1.359(0.452)

known. In the practice used in this paper, the instrument set is identical to the original regressors, X, apart from where the endogenous variable has been replaced with an instrument. In specially written macros we combine the IV and CMS methods to provide consistent estimation and also enable model comparisons.

2   CMS and IV Estimation with Simulated Data

Fifty datasets were simulated, each consisting of 36 groups of pupils, each group (or school) containing a number of pupils varying between 11 and 33. Random $N(0,1)$ components for unmeasured heterogeneity were generated for schools (lvs_i) and pupils (lvp_{ij}) and summed to form a latent variable (lv_{ij}). We generated $X_{ij} = Cons + (lv_{ij}/2) + N(0,1)\ error$ as a prior test score. In operation X_{ij} was centred to give C_{ij} used below. Then to form an instrument for X and correlated with it but independent of the latent variable, the variable $I_{ij} = X$ -$(lv_{ij}/2)$ was created. A 'current test score' was then formed by $y_{ij} = \beta C_{ij} + lvs_i + lvp_{ij} + e_{ij}$, with e_{ij} a further generated $N(0,1)$ error. This model (appropriately) includes random components also used in forming C_{ij}, to make the latter endogenous. The "test score" is then divided into 11 'observed' categories by the evenly spaced 'cut points' in Table 1. Models were fitted for each data set using a probit link adaptation of MULTICAT macros of MlwiN. This probit version is obtainable by e-mail from the authors.

Table 1 gives parameter values (and standard errors) for combinations of useage of CMS and IV. For neither used it is not surprising to find non recovery of parameter values. Method 2 re-scales but poor recovery may be due to ignoring endogeneity. Method 3 is beset by the original problems of scaling. However, the method using both CMS and IV estimation performs

quite successfully though school variance is over-estimated. A common objection to IV methods is their imprecision. However, it should be noted that standard errors here are quite respectable.

It may be noted that application of the CMS method here differs and improves on that of Fielding (2003) for this situation. The rescaling is done iteratively within the second model. So that the results of using methods 2 and 4 can be compared with the parameter values used to create the simulated data, the known parameter values for the null model have been used rather than estimates.

3   Estimation with Data from Birmingham Schools

The data arise from 4421 children aged around 7 years in 114 schools in Birmingham, UK. GENDER is a male dummy, FSM is a dummy for school meal eligibility. Ethnic background first language overlap and may confound so compound categories were formed giving rise to 14 dummies (AMCLANG1-14). CTRDAGE was age in months centred on 84. Two school context variables were used: the % of pupils with FSM=1 (PCTFSM) and average % of baseline assessments that were graded above 2 (AVPCT-BASEGT2). Baseline assessments of ability carried out by teachers at the beginning of the school year in four areas of mathematics (number, algebra, shape and space, handling data) and three areas of language and literacy (speaking and listening, reading, writing) with pupils being given a grade of (in descending order) 3, 2, 1, 0 in each of the seven areas . Towards the end of the school year, the pupils took the Key Stage 1 Mathematics Standard Assessment Task. Pupils were given grades from this of (in descending order) 3, 2a, 2b, 2c, 1, 0. We use this variable, having six ordered categories, as the response variable in the modelling. Fielding (1999) gives fuller details.

An initial null model (model A) gave the basis for scale factor adjustment. Following the example and reasons of Fielding (1999), the four models detailed in Table 2 were fitted. For these models how were the instruments for endogenous baseline tests formed? From experimentation, it is apparent some available variables are not related to the response. These are whether or not a pupil attended at least one full term of nursery school and 10 of the 14 AMCLANG variables. Instruments could be formed as predictions from fixed part of a multilevel model of baseline assessment variables using these 11 variables are regressors. However, a complication arises here since there are seven endogenous baselines. The efficiency of IV estimation is affected by the canonical correlations between the set of endogenous variables and the set of instrumental variables. Loose correlations of each baseline with its instrument means that canonical correlations and thus efficiency of estimates will be low. To overcome this, we used the first principal component of the seven baselines (59.9% of variation) was used as a regressor in the model and a parallel instrument formed for it.

TABLE 2. Parameter estimates and standard errors with use of CMS and IV estimation

Coefficient	Model B	Model C	Model D	Model E
Cut-point 1	-2.013(0.051)	-2.225(0.061)	-2.144(0.067)	-2.306(0.089)
Cut-point 2	-0.848(0.051)	-1.102(0.052)	-0.998(0.067)	-1.218(0.089)
Cut-point 3	-0.202(0.051)	-0.473(0.050)	-0.361(0.067)	-0.616(0.089)
Cut-point 4	0.331(0.051)	0.044(0.050)	0.164(0.067)	-0.120(0.089)
Cut-point 5	1.056(0.051)	0.742(0.051)	0.880(0.067)	0.556(0.089)
1st PC for baseline tests	0.247(0.058)		0.229(0.056)	0.185(0.050)
GENDER		0.007(0.030)	-0.085(0.024)	-0.070(0.021)
FSM		0.309(0.033)	0.291(0.047)	0.193(0.021)
CTRDAGE		-0.061(0.004)	-0.033(0.008)	-0.036(0.007)
AMCLANG2		0.053(0.062)	0.138(0.070)	0.103(0.058)
AMCLANG11		-0.629(0.336)	-0.707(0.142)	-0.639(0.118)
AMCLANG12		-0.765(0.436)	-1.201(0.208)	-1.143(0.171)
PCTFSM				0.006(0.002)
AVPCTBASEGT2				0.030(0.018)
School variance	0.236(0.037)	0.176(0.027)	0.201(0.031)	0.141(0.023)

Results from fitting these models are shown in table 2. AMCLANG2 corresponds to an Afro-Caribbean ethnic background with first language English; AMCLANG11 corresponds to a Chinese ethnic background with first language not English; AMCLANG12 corresponds to a Vietnamese ethnic background with first language not English. All are relative to a White ethnic background with first language English.

It should be noted importantly that, as with the results of the simulations, the standard errors of the estimates obtained are respectable for all models including B, D and E where IV estimation takes place. Unlike many other published applications, in estimation we have also accounted for the possible endogeneity of baseline variables.

4   Discussion

The simulation analysis indicates that both CMS and IV estimation are necessary for model comparisons can both be applied successfully. They have also been successfully applied to the dataset. In further investigation , not reported here we have also not used IV and the estimated effects are very different. The MLwiN macro files used are available online: www.herts.ac.uk/business/staff_public/nhspencer_public/research.

References

Fielding, A. (1999) Why use arbitrary points scores: ordered categories in models of educational progress. *Journal of the Royal Statistical Society, Series A*, 162, 3, pp 303-328.

Fielding, A. (2003) Scaling for residual variance components of ordered category responses in generalised linear mixed multilevel models. *Quality and Quantity* (in press).

Fielding, A. (2002) Ordered category responses and random effects in multilevel and other complex structures: scored and generalised linear models *in Multilevel Modeling: Methodological Advances, Issues and Applications* (S. Reise & N. Duan, eds.), Erlbaum: New Jersey.

Snijders, T.A.B. and Bosker, R.J. (1999) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, Sage Publications: London.

Spencer, N. H. (2002) Combining modelling strategies to analyse teaching styles data. *Quality and Quantity*, 36, 2, pp 113-127.

Spencer, N. H. & Fielding, A. (2002) A comparison of modelling strategies for value-added analyses of educational data. *Computational Statistics*, 17, 1, pp 103-116.

Bayesian techniques for modelling volcanic processes

Claudia Furlan[1]

[1] Dipartimento di Scienze Statistiche, Via C. Battisti 241/243, 35121 Padova, Italia.

Abstract: Extreme value theory is the branch of statistics inferring extreme events in random processes. Bayesian estimation in this field offers many advantages. We use techniques from extreme value theory to estimate by Bayesian methods the probability distribution of extreme volcanic eruptions that are subject to a historical recording bias.

Keywords: Extreme values; Bayesian techniques; censored data; volcano eruptions.

1 Introduction

Elsewhere in these proceedings, Coles (2004) discusses a censored point process model to describe extreme volcanic eruptions, with inference based on maximum likelihood. Moreover there are limitations in this approach to inference and Bayesian techniques offer an alternative that is often preferable. There are number of reasons why a Bayesian analysis of extreme value data might be desirable. First, owing to scarcity of data, there is the facility to include information through a prior distribution. Second, the output of a Bayesian analysis – the posterior distribution – provides a more complete inference than the corresponding maximum likelihood analysis. In particular, since the objective of an extreme value analysis is usually an estimate of the probability of future events reaching extreme levels, expression through predictive distribution is natural. Third, Markov chain Monte Carlo techniques allow to estimated more complex parameter structure and also when the parameter dimension in unknown. In the volcano setting, we will be able to work with more flexible model structures.

2 Historical catalogue of volcanic eruptions

The data represented in Figure 1 have been recorded in a historical catalogue over the past two millennia.
The magnitude is defined by $M = \log(m) - 7$, where m is the erupted mass in Kg. The structure of these data suggests an extreme value analysis
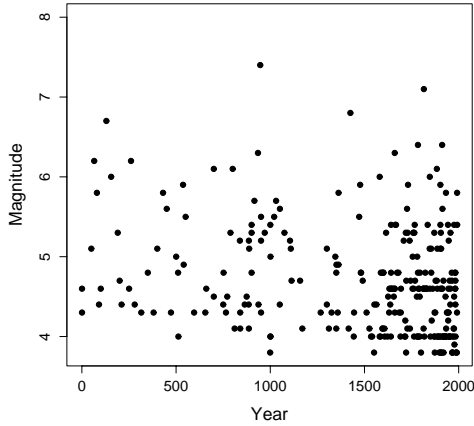
FIGURE 1. Volcanic eruptions exceeding 3.7 M.

(Coles, 2001), but the process does not seem stationary. Indeed, looking at points below 5M, the rate of volcanic activity seems much greater in recent years, while above 6M the rate seems more or less uniform throughout. This suggests that, for relatively small events, there was a difficulty in recording volcanic events especially further back in time. In Coles (2004), the events (t_i, x_i), with t_i being time re-scaled to $[0,1]$ and x_i denoting the magnitude, were modeled with a Poisson process over a threshold with intensity:

$$\lambda_M(t, x) = p(t, x)\lambda(t, x) \tag{1}$$

where

$$\lambda(t, x) = \frac{1}{\sigma} \left[1 + \xi \frac{(x - \mu)}{\sigma} \right]_+^{-1/\xi - 1} \tag{2}$$

with $\sigma > 0$ and $a_+ = \max(a, 0)$, and with a constrained parametric model for $p(t, x)$. Component (2) is based on standard extreme values arguments whereas (1) summarizes our belief about the recording mechanism. In this article, guided by Figure 1, we consider, as an alternative, a changepoint specification for $p(t, x)$.

2.1   A Changepoint Model

Looking again at Figure 1, the process looks stationary, at least to the eye, over the last 500 years. An alternative formulation for $p(t, x)$ might therefore be in terms of a changepoint model.

TABLE 1. Posterior expected values of μ, σ, ξ, a, b and mode of the posterior distribution of k.

$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	$\hat{a}$	$\hat{b}$	k (mode)
2.52	1.42	-0.25	-3.24	0.38	1587
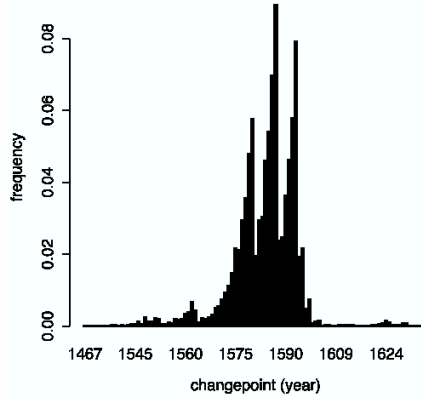
FIGURE 2. Posterior distribution of k, referred to model (3).

Specifically, if $\lambda_M(t,x) = p(t,x)\lambda(t,x)$ is the density of the Point Process model, a viable censoring function is:

$$p(t,x) = \begin{cases} \frac{\exp(a+bx)}{1+\exp(a+bx)} & t \leq k \\ 1 & t > k, \end{cases} \qquad (3)$$

for some $k \in [0, 2000]$ (scaled back to years). Provided $b > 0$, this ensures that $p(t,x) \uparrow 1$ as $x \uparrow \infty$.

To estimate the parameters $\theta = (\mu, \sigma, \xi, a, b, k)$ we have used Markov Chain Monte Carlo techniques, with a Metropolis-Hastings algorithm (Gilks *et al.*, 1996). See Table 1 for a summary of the posterior expected values of μ, σ, ξ, a, b and the mode of the posterior distribution of k, and to Figure 2 for a graphical representation of the posterior distribution of k. The estimates of μ, σ, ξ are broadly consistent with those of Coles (2004), while the estimate of k seems in accord with the visual impression of Figure 1.

Though Figure 1 suggests the presence of just one changepoint, we also tried to estimate a model with an arbitrary number of changepoints (generalizing (3)), introducing a parameter space with unknown dimension. Let N_k be the number of changepoints and $K = (k_1, \ldots, k_{N_k})$ the vector

of changepoints; also set $k_0 = 0$ and $k_\infty = 2000$. Then, we define:

$$p(t,x) = \begin{cases} \frac{\exp(a_i+b_ix)}{1+\exp(a_i+b_ix)} & (k_i, k_{i+1}) \quad i = 0, \ldots, N_k - 1 \\ 1 & (k_{N_k}, K_\infty). \end{cases} \quad (4)$$

The technique used is Reversible Jump Markov Chain Monte Carlo (Green, 1995). Despite the extra flexibility, the inference points very strongly to the presence of a single changepoint only.

2.2 Predictive distribution

Prediction is also handled better within a Bayesian setting. If z denotes a future volcanic eruption having probability distribution function $G(z|\theta)$ and $f(\theta|x)$ is the posterior distribution of θ on the basis of observed volcano eruptions x, then:

$$\Pr\{Z \leq z|x\} = \int_\Theta G(z|\theta)f(\theta|x)d\theta \quad (5)$$

is the predictive distribution of z given x. Compared with other approaches to prediction, the predictive distribution has the advantage that it reflects uncertainty in the model –the $f(\theta|x)$ term – and uncertainty due to the variability in future observations – the $G(z|\theta)$ term. Whilst the predictive distribution may seem intractable, it is easily approximated if the posterior distribution has itself been estimated by simulation, using for example MCMC. After deletion of the values generated in the settling-in period, the procedure leads to a sample $\theta_1, \ldots, \theta_s$ that may be regarded as observations from the stationary distribution $f(\theta|x)$ and

$$\Pr\{Z > z|x\} \approx \frac{1}{s}\sum_{i=1}^{s}(1 - G(z|\theta_i)) = \frac{1}{s}\sum_{i=1}^{s}[1 + \xi(z-u)/\tilde{\sigma}]_+^{-1/\xi} \quad (6)$$

Based on the changepoint model, a graphical representation of the estimated predictive distribution of volcanic magnitude conditional on an exceedance of 4M, is shown in Figure 3.

3   Conclusions

Reformulating the basic censored point process model of Coles (2004) within a Bayesian framework leads to several advantages. Here we have considered two: the specification of a changepoint model for the censoring mechanism and the calculation of the predictive distribution of extreme volcanic magnitudes. Both aspects give a preferential interpretation relative to a classic inference.
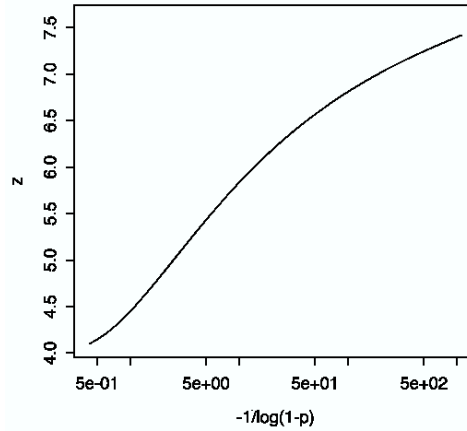
FIGURE 3. Predictive conditional distribution of $p = P(Z > z | x > 4)$ versus z, on standard extreme value scale.

References

Coles, S.G. (2001). *An introduction to statistical modeling of extreme values.* London: Springer.

Coles, S.G. (2004) A Censored Point Process Model for Extreme Volcanic Eruptions, *In Proceeding.*

Gilks, W.R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice.* London: Chapman & Hall.

Green, P. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, **82**, 711-732.

Bayesian analysis of transmission dynamics of experimental epidemics

George Streftaris[1] and Gavin J Gibson[1]

[1] Actuarial Mathematics and Statistics, School of Mathematical and Computer
Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK

Abstract: We analyse data from two foot-and-mouth disease experiments for which previous studies have indicated lower levels of virus in the blood of sheep infected in the later stages of the epidemic. By using a non-Markovian stochastic compartmental model in a Bayesian approach, coupled with Markov chain Monte Carlo techniques, we are able to relax earlier assumptions regarding possible pathways of infection, and to use the data to reconstruct the infectious network. Thus, the complex interactions among level of viraemia, individual infectiousness and temporal position in the epidemic process can be investigated.

1   Introduction

We investigate the transmission dynamics of a certain type of foot-and-mouth disease (FMD) virus under experimental conditions, using an SEIR (Susceptible-Exposed-Infectious-Removed) non-Markovian compartmental model for partially observed epidemic processes. Previous analyses of experimental data from FMD outbreaks in non-homogeneously mixing populations of sheep have suggested a decline of viraemic level in animals infected in the later stages of the epidemic. However, these studies do not take into account possible variation in the length of the chain of virus transmission for each animal, which is implicit in the non-observed transmission process. We employ powerful Markov chain Monte Carlo (MCMC) methods (e.g. Tierney, 1994) for statistical inference, to address epidemiological issues under a Bayesian framework that accounts for all available information and associated uncertainty in a coherent approach. Such methodology is being increasingly employed for inference in stochastic compartmental epidemic models (Gibson and Renshaw, 1998; O'Neill and Roberts, 1999). The analysis provides estimation of epidemiological parameters, and also allows the investigation of more complex characteristics of the virus transmission process, relying on stochastic realisations of the unobserved network of infectious contacts.

Data were collected during two experiments (Hughes *et al.*, 2002), in which 32 sheep were randomly allocated to four groups (G1 to G4), and the first group animals were inoculated with the same FMD virus dose. The virus

was then passed to animals of the remaining groups, through a process designed so that throughout the duration of the experiment each group spent 24 hours mixing with a given group of 'donor' animals, followed by 24 hours in the presence of a given 'recipient' group. Viraemic diagnosis was based on daily blood samples. The data used in this paper consist of individual records of the day of onset and cessation of viraemia and the peak viraemic levels.

Our analysis aims at addressing three main issues: the quantification of basic disease transmission characteristics, such as the contact rate and the duration of latent periods; the study of the relation between level of viraemia and infectiousness; and the investigation of a hypothesis that infectiousness declines along the chain of virus transmission.

2   Model and methodology

We represent the spread of the epidemic through an SEIR model (Bailey, 1975) and following the work in Streftaris and Gibson (2004a) we employ the two-parameter Weibull(ν, λ) distribution to describe sojourn times in various compartments. We use n to denote the number of viraemic animals in the population. The observation period of the epidemic is represented in our model by the time interval $[0, T]$, defining its start as the inoculation time and its end as the time of the last recorded event (last recovery). The design of the experiments mimics a non-homogeneous population mixing pattern, according to which the groups mix in pairs on alternate days. If $\boldsymbol{\theta} = (\alpha, \beta, \gamma_1, \delta_1, \gamma_2, \delta_2, \nu, \lambda)^{\mathrm{T}}$ denotes the vector of model parameters, the likelihood of the complete data (assuming perfect observation of the epidemic) can be written as

$$L(\boldsymbol{\theta}; \mathbf{e}, \mathbf{s}, \mathbf{r}) = \prod_{j \in \mathcal{E}} \left[\beta \sum_{l=1}^{n} \{ v_l^\alpha i_l(G_j, e_j) \} \right] \times \exp \left\{ - \int_0^T \beta C(t) dt \right\}$$

$$\times \prod_{j \in \mathcal{I}_1} f_1(s_j - e_j; \gamma_1, \delta_1) \times \prod_{j \in \mathcal{I}_{2,3,4}} f_2(s_j - e_j; \gamma_2, \delta_2) \times \prod_{j \in \mathcal{R}} f_3(r_j - s_j; \nu, \lambda),$$

with β denoting the rate of infection per possible susceptible-infectious contact weighed by the associated infectivity; e_j, s_j, r_j denote respectively the time of exposure, start of infectious period and recovery of animal j, and $\mathbf{e}, \mathbf{s}, \mathbf{r}$ are the corresponding vectors; G_j is the group to which animal j belongs; $f_1(\cdot), f_2(\cdot)$ denote the Weibull densities for the latent periods of animals in G1 and G2-G4 respectively, and $f_3(\cdot)$ is the Weibull density of the infectious period. We consider the peak viraemic level of each infectious sheep as a potential factor affecting the infective challenge exerted on each susceptible animal. The possible influence is modelled as the sum of a power function of the individual viraemic levels $v_l, l = 1, \ldots, n$, allowing the power level, denoted by α, to be estimated as a model parameter.

The function $i_l(k, t)$ provides an indicator factor such that for $l = 1, \ldots, n$, $i_l(k, t) = 1$ if at time t animal l is infectious and mixing only with group k, or zero otherwise. Also, $\mathcal{E}, \mathcal{I}_1, \mathcal{I}_{2,3,4}, \mathcal{R}$ denote the sets of exposed (G2-G4), infectious (G1, G2-G4) and recovered animals at the end of the experiment, while $\beta C(t)$ represents the total infective force on the susceptible population at time t, given the mixing pattern and the infectious state of the population at that time (see Streftaris and Gibson, 2004b).

The available information in the likelihood is only partial, as the exposure times for naturally infected animals, $e_l, l \in \mathcal{E}$, are not known, and the recorded times of infectiousness onset (s_l) and recovery (r_l) correspond to sampling carried out every 24 hours, and are therefore not exact. For reliable inferences the hidden aspects of the epidemic process must be accounted for and any associated uncertainty should be appropriately addressed.

2.1   Bayesian investigation of hidden infection process

We follow a Bayesian approach, under which the unobserved events in the transmission process of the disease are represented as nuisance parameters. Assuming independent gamma prior distributions for all model parameters, the joint posterior density $p(\boldsymbol{\theta}|\mathbf{e}, \mathbf{s}, \mathbf{r}) \propto L(\boldsymbol{\theta}; \mathbf{e}, \mathbf{s}, \mathbf{r})\pi(\boldsymbol{\theta})$, is investigated and inferences on model parameters are extracted from the respective marginal densities. The joint posterior density is given in an analytically intractable form, and therefore inference will rely on computationally intensive estimation methods. We use a MCMC algorithm that comprises a combination of Gibbs sampling, independence Metropolis–Hastings and random-walk Metropolis steps, in a manner similar to that described in Streftaris and Gibson (2004a).

To investigate the effect of the length of the infection chain to the detected level of viraemia we first consider stochastic reconstructions of the network of infectious contacts, within our MCMC scheme. Possible infectious pathways can be determined via the posterior distribution of the unobserved times of exposure to the disease, by linking each viral exposure to an available infectious individual, using a probability weighted by the individual's infectiousness. Thus, the length of the infection chain for each animal is determined, providing a partition of the population to infection generation categories. We assess the possible effect on the exhibited viraemia using ANOVA to test a null hypothesis of no differences in viraemic levels along the increasing length of the infection chain, obtaining an associated p-value for the null hypothesis. The whole posterior distribution of these p-values can then be obtained based on data from the MCMC output (cf. Meng, 1994).

3   Results

Posterior estimates of the parameters quantifying the spread of the FMD in the two studied experiments are obtained. Characteristics of interest are: the transmission (or contact) parameter β; the duration of the latent (incubation) period of the disease; and the parameter α used to assess a possible relation between blood viral load and infectiousness of individual sheep. The corresponding posterior densities are shown in Figure 1. The mean latent period appears to be shorter than usually reported in the literature (especially for G2-G4 animals), reflecting the highly intensive infection process in the experiments. The posterior densities of all model parameters are consistent with the assumption of the same underlying epidemic process in the two experimental occasions. Under the assumption of a non-informative $Ga(1, 0.001)$ prior distribution for parameter α, its posterior distribution indicates that the information in the data supports non-zero values of the parameter. Our analysis therefore suggests that individual blood viral load affects the infective challenge exerted on susceptible animals in both experiments. The results also reveal a possible decline in viraemia in one of the two experimental outbreaks, as the corresponding posterior distribution favours small p-values (Streftaris and Gibson, 2004b).
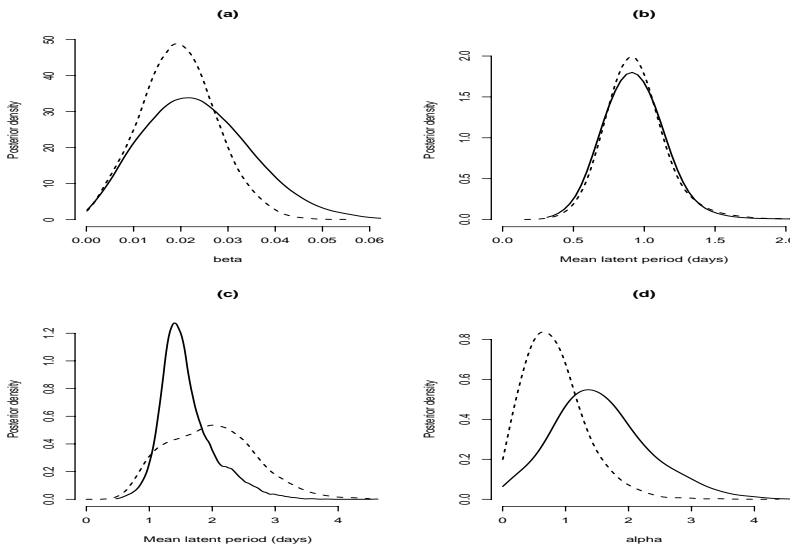
FIGURE 1. Posterior densities of the characteristics of the transmission of FMD virus in sheep under experimental conditions. (a) β; (b) Mean latent period G1; (c) Mean latent period G2-G4; (d) α. The solid and dashed lines correspond to Experiment 1 and Experiment 2 respectively.

4   Discussion

The power of the modelling and methodology used in this paper to address the question of possible relations among level of viraemia, infectiousness and length of infection chain, was assessed through a simulation study. Epidemic data were generated under various scenarios assuming appropriate combinations of the effect of viraemia on infectiousness ($\alpha = 0$ or 1), and decreasing or unchanged levels of viraemia. In all cases our analysis was able to correctly identify the presence (or not) of both effects.

Assessment of the fit of the model with the use of Bayesian latent residuals, has suggested a possible under-dispersion of the unobserved times of infectious contacts. An assumption of gamma distributed tolerance levels to the disease may then be incorporated in the model. This issue, together with others related to alternative distributions for sojourn times, leads to a question of model choice for partially observed epidemics, which we are currently addressing using simulation studies and Bayes factors related methodology.

References

Bailey, N.T.J. (1975). *The Mathematical Theory of Infectious Diseases and its Applications, 2nd ed.* London: Griffin.

Gibson, G.J. and Renshaw, E. (1998). Estimating parameters in stochastic compartmental models using Markov chain Monte Carlo methods. *IMA J. Math. Appl. Med. Biol.* **15**, 19–40.

Hughes, G.H., Mioulet, V., Haydon, D.T., Kitching, R.P., Donaldson, A.I. and Woolhouse, M.E.J. (2002). Serial passage of foot-and-mouth disease virus in sheep reveals declining levels of viraemia over time. *J. Gen. Virol.* **83**, 1907–1914.

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

O'Neill, P.D. and Roberts, G.O. (1999). Bayesian inference for partially observed stochastic epidemics. *J. R. Statist. Soc. A* **162**, 121–129.

Streftaris, G. & Gibson, G.J. (2004*a*). Bayesian inference for stochastic epidemics in closed populations. *Statist. Modelling* **4**, 63–75.

Streftaris, G. & Gibson, G.J. (2004*b*). Bayesian analysis of experimental epidemics of foot-and-mouth disease. *Proc. R. Soc. Lond.* B (In press).

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22**, 1701–1762.

Weighted Estimation of Variance Components and Fixed Effects in Small Area Models

A.F. Militino[1], M.D. Ugarte[1] and T. Goicoa[1]

[1] Departamento de Estadística e Investigación Operativa, Universidad Pública de Navarra, Campus de Arrosadía, 31006 Pamplona, Spain
e-mail: militino@unavarra.es

Abstract: The aim of this paper is to propose a unit level linear mixed model and an area level linear mixed model where both, the variance components and the coefficients of the model are estimated using weights. The models performance is illustrated by estimating the total area occupied by olive trees in a region called Comarca IV, located in Navarra, Spain. Small area linear mixed models have been used for similar purposes using regular quadrats (also called segments) as sampling units, and assuming that these are fully included in the study domain. However when this does not happen, the sampling units are very different in size, leading to an extra variability within areas. Then, the inclusion of weights in the model is recommended.

Keywords: Borrow information; Linear mixed models; Variance components.

1   Introduction

There is an increasing demand in local and central Governments in knowing precise estimates in domains where the size of the samples is small or even zero. These domains are called small areas. Traditionally, the sample sizes are chosen to provide reliable estimates for large geographical regions or aggregates of small areas. However, the statistical methods used for large domains can rarely be applied to small ones. Then, the problem of small area estimation is twofold. First, the fundamental question of producing reliable estimates of characteristics of interest and second, the assessment of the estimation error. When the sample in a given area is very small, a solution to the estimation problem is to *borrow strength* from related areas by means of auxiliary information. Different model-based methods to accomplish small area issues have been proposed in the literature (for a good review see Rao, 2003). Battese, Harter and Fuller (1988), popularized the use of linear mixed models in agricultural small area problems. They gave a prediction of the mean hectares of soybeans and corn per segment in 12 counties of Iowa with 36 segments, using as auxiliary information the classified corn and soybean hectares provided by satellite images. The authors

consider a simple random sampling plan and segments of 250 hectares entirely included in the study domain. A common approach to account for other sampling plans by using sampling weights has been done by Prasad and Rao (1999) and You and Rao (2002) who develop design-consistent small area estimation models. In these models the variance components are estimated from a unit-level model but the authors do not incorporate weights into the estimation process. Then, more difficulties arise to validate the model.

The aim of this paper is to propose a unit level linear mixed model and an area level linear mixed model where both, the variance components and the coefficients of the model are estimated using weights. The models performance is illustrated by estimating the total area of olive trees in a region called Comarca IV, located in the central part of Navarra, Spain. The olive oil industry is becoming very important and there is a general interest in determining the land area occupied by this crop in different regions mainly for two reasons: to control the olive-oil production, and to distribute European financial help. Traditionally, small area linear mixed models have been used for similar purposes based on the common definition of regular quadrats (also called segments) as sampling units, and assuming that these are fully included in the study domain. However, one important feature of this sample is that the square segments are very small, only of 4 hectares, and often, not completely included in the very irregular study domain. The size of sampled segments was limited by the precision of satellite images and could not be reduced. Figure 1 shows the big irregularity of the many spots that constitute the study domain and how the majority of sampled segments are scarcely included there.

2 Weighted Linear Mixed Models

Battese, Harter and Fuller (1988), explained the reported hectares of soybeans or corn in the sample segments within counties as a function of the satellite data for those sample segments, such that the reported hectares are positively correlated within given counties but uncorrelated from different counties. The model is given by

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + u_{ij}, \quad i = 1, \ldots, t, \quad j = 1, \ldots, n_i, \quad (1)$$

where in the ith county ($i = 1, \ldots, t$), y_{ij} is the number of hectares of soybean (or corn) in the jth segment, n_i is the number of sampled segments, x_{ij1} and x_{ij2} are the jth classified hectares of soybeans and corn respectively, and β_0, β_1 and β_2 are unknown parameters. The random error u_{ij} associated with the reported area y_{ij} is expressed by

$$u_{ij} = v_i + e_{ij}, \quad v_i \sim N(0, \sigma_v^2), \quad \text{and} \quad e_{ij} \sim N(0, \sigma_e^2), \quad (2)$$
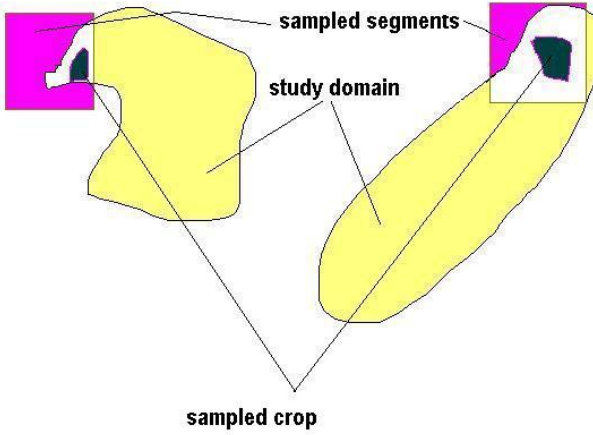
FIGURE 1. Study domain and sampled crops in 4 ha. segments.

where v_i is the ith county effect and e_{ij} is the random error associated with the jth sample segment within the ith county. The random effects v_i are assumed to be independent of the random errors e_{ij} ($j = 1, \ldots, n_i; i = 1, \ldots, t$). These authors do not include weights in the estimation process. To account for heteroscedasticity within small areas, we propose the use of weights to estimate both, variance components and fixed effects. The proposed model is a weighted unit level linear mixed model, where the auxiliary information is available for every sampled unit and for the whole area. It is given by

$$y_{ij} = \mathbf{x}'_{ij}\beta + v_i + \epsilon_{ij}, \quad v_i \sim N(0, \sigma_v^2), \quad \epsilon_{ij} \sim N(0, \sigma_e^2/w_{ij}), \qquad (3)$$

where in the ith county ($i = 1, \ldots, t$), y_{ij} is the number of hectares of crop in the jth segment, n_i is the number of sampled segments, $\mathbf{x}'_{ij}$ is the jth classified hectares of crop and w_{ij} are the weights. The predictor of the ith-mean is given by

$$\tilde{\bar{y}}_{iw} = \bar{\mathbf{x}}'_{i(p)}\tilde{\beta}_w + \gamma_{iw}(\bar{y}_{iw} - \bar{\mathbf{x}}'_{iw}\tilde{\beta}_w), \quad i = 1, \ldots, t, \qquad (4)$$

and it is estimated by

$$\hat{\tilde{\bar{y}}}_{iw} = \bar{\mathbf{x}}'_{i(p)}\hat{\beta}_w + \hat{\gamma}_{iw}(\bar{y}_{iw} - \bar{\mathbf{x}}'_{iw}\hat{\beta}_w), \quad i = 1, \ldots, t, \qquad (5)$$

where $\bar{y}_{iw} = \sum_{j=1}^{n_i} w_{ij}y_{ij}/w_{i\cdot}$, $\bar{\mathbf{x}}_{iw} = \sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}/w_{i\cdot}$, $\tilde{\beta}$ is the weighted least squares estimator of β assuming that the variance components σ_e^2 and σ_v^2 are known and $\hat{\beta}_w = \tilde{\beta}_w(\hat{\sigma}_e^2, \hat{\sigma}_v^2)$ is the estimate of $\tilde{\beta}$ after estimating the variance components. $\hat{\gamma}_{iw}$ is the plug-in estimator of $\gamma_{iw} = \sigma_v^2/(\sigma_v^2 + \sigma_e^2/w_{i\cdot})$ and $\bar{x}_{i(p)}$ is the population mean of the auxiliary variable. The predictor in Equation (4) depends on the variance components $\sigma^2 = (\sigma_v^2, \sigma_e^2)$, but in practice, they are unknown. A common way of estimating the variance components is by using the fitting of constants or moments method (Searle, Casella and McCullogh, 1992), that yields unbiased estimators without depending on normality assumptions. The estimators have closed expressions and are easy to compute. This method is used by You and Rao (2002), but they do not include weights into the estimation procedure. In this paper we modify this technique by including weights into the variance component estimation process. The mean squared error of the prediction is also re-estimated following the approximation proposed by Prasad and Rao (1990). The models validation is also presented.

3   Conclusions

When the variability within small areas is very different and heteroscedasticity is present, the use of weights is specially recommended. In the particular application considered here, we show how the heteroscedasticity is better corrected in models including weights into the variance component estimation process, both in unit level models and in area level models. We illustrate the results with the estimation of total land area occupied by olive trees in a particular region of Navarra, Spain. The data consists of 49 segments of 4 hectares drawn by simple random sampling in 8 non-irrigated areas. We estimate the total number of hectares and their corresponding mean squared prediction error in each small area using the models that we propose in this paper. A comparison is done with other models already proposed in the literature.

References

Battese, G. E., Harter, R. M. and Fuller, W. A. (1988). An Error Components Model for Prediction of Country Crop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association*, **83**, 28-36.

Prasad, N.G.N. and Rao, J.N.K. (1990). The Estimation of Mean Squared Error of Small Area Estimators. *Journal of the American Statistical Association*, **85**, 163-171.

Prasad, N.G.N. and Rao, J.N.K. (1999). On Robust Small Area Estimation Using a Simple Random Effects Model. *Survey Methodology*, **25**, 67-72.

Rao, J.N.K. (2003). *Small Area Estimation*. Wiley Series in Survey Methodology.

Searle, S.R., Casella, G., McCullogh, C.E. (1992). *Variance Components*. Wiley Series in Probability and Statistics.

You, Y. and Rao J.N.K. (2002). A pseudo-empirical best linear prediction approach to small-area estimation using survey weights. *Canadian Journal of Statistics*, **30**, 431-439.

A polytomous response multilevel model with a non ignorable selection mechanism

Leonardo Grilli[1] and Carla Rampichini[1]

[1] Dipartimento di Statistica "G. Parenti" Viale Morgagni 59 50134 Firenze

Abstract: The aim of the paper is to specify and fit a multilevel model for a polytomous response in presence of potential selection bias. The work is motivated by the analysis of the way of acquisition of the skills of university graduates. In order to taking into account the features of the data, a suitable multivariate multilevel model for polytomous responses with a non-ignorable missing data mechanism is developed and fitted by means of maximum likelihood with adaptive Gaussian quadrature. In the application the multilevel structure has a crucial role, while selection bias results negligible.

Keywords: multilevel models; polytomous response; selection bias; university evaluation.

1 Introduction

Selection bias may arise when the selection mechanism depends on unobserved variables correlated with the error terms of the statistical model of interest. A classical way to correct the selection bias (Heckman, 1979) is to add an equation which explicitly models the selection mechanism. Applications of this approach in the multilevel framework are still rare (e.g. Borgoni and Billari, 2002) and, as far as we know, none of them concerns the polytomous case.

The paper was motivated by the analysis of data gathered from a telephone survey conducted, about two years after the degree, on the 2000's graduates of the University of Florence. Particularly, interest is in the analysis of some skills which may be requested for the current job. The analysis of such data raises several methodological issues: (a) the response is composed by a set of categorical variables, with potential selection bias due to the design of the questionnaire: for each skill a first question asks if the graduate currently uses it, while, in case of an affirmative response, a second question asks where the skill was acquired, so for all the graduates that do not use the skill the second question is missing, causing a potential selection bias; (b) for each skill, the second question has a polytomous response, aggregated to three categories: the skill was acquired during the degree programme, at workplace or otherwise; (c) the data have a hierarchical structure (items within graduates and graduates within degree programmes), so that the

observations are correlated. The questionnaire includes eight skills, and for each skill two questions are asked. In the present work each skill is analyzed separately.

2   The model

In the case of a polytomous response with M categories (alternatives), the model with selection has M equations, one for the dichotomous selection indicator (e.g. current use of the skill) and $(M-1)$ for the polytomous response of interest (e.g. way of acquisition of the skill), where the probability of the reference alternative ($m = 1$) is obtained by difference. Indexing the cluster (e.g. degree programme) by $j = 1, 2, \cdots, J$ and the subject of the j-th cluster (e.g. graduate) by $i = 1, 2, \cdots, n_j$, and assuming a logit link for both sets of equations, the model is:

$$P(Y_{ij}^S = 1 \mid \boldsymbol{x}_{ij}^S, \xi_j^S, \delta_{ij}^S) = \frac{\exp\{\alpha^S + \boldsymbol{\beta}^{S\,\prime}\boldsymbol{x}_{ij}^S + \xi_j^S + \delta_{ij}^S\}}{1 + \exp\{\alpha^S + \boldsymbol{\beta}^{S\,\prime}\boldsymbol{x}_{ij}^S + \xi_j^S + \delta_{ij}^S\}} \quad (1)$$

$$P(Y_{ij}^P = m \mid \boldsymbol{x}_{ij}^P, \boldsymbol{\xi}_j^P, \boldsymbol{\delta}_{ij}^P) = \frac{\exp\{\eta_{ij}^{P(m)}\}}{1 + \sum_{l=2}^{M} \exp\{\eta_{ij}^{P(l)}\}} \quad (m = 2, \cdots, M)$$

where the variable Y_{ij}^P is observed *if and only if* $Y_{ij}^S = 1$. Moreover $\boldsymbol{\xi}_j^P = (\xi_j^{P(2)}, \ldots, \xi_j^{P(M)})'$ and $\boldsymbol{\delta}_{ij}^P = (\delta_{ij}^{P(2)}, \ldots, \delta_{ij}^{P(M)})'$. The linear predictor of the m-th alternative is $\eta_{ij}^{P(m)} = \alpha^{P(m)} + \boldsymbol{\beta}^{P(m)\,\prime}\boldsymbol{x}_{ij}^P + \xi_j^{P(m)} + \delta_{ij}^{P(m)}$. The superscript S denotes the variables and parameters of the selection equation, while the superscript P denotes the variables and parameters of the principal (polytomous) equations; in particular $P(m)$ refers to the m-th alternative. The S and P sets of equations may have distinct covariates, $\boldsymbol{x}_{ij}^S$ and $\boldsymbol{x}_{ij}^P$, though there are no alternative specific covariates in the present specification of the polytomous model; moreover each equation has different parameters: α^S and $\boldsymbol{\beta}^S$ for the selection equation, and $\alpha^{P(m)}$ and $\boldsymbol{\beta}^{P(m)}$ ($m = 2, \cdots, M$) for the principal equations, where the superscript $P(m)$ indicates that the parameters vary with the alternative. The ξ_js and δ_{ij}s are random variables representing unobserved heterogeneity at cluster and subject level, respectively, with the following distributional assumptions: errors at different levels are independent; the random vector $(\xi_j^S, \xi_j^{P(2)}, \cdots, \xi_j^{P(M)})'$ has a multivariate normal distribution, with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{\Sigma}_\xi$; while the random vector $(\delta_{ij}^S, \delta_{ij}^{P(2)}, \cdots, \delta_{ij}^{P(M)})'$ has a multivariate normal distribution, with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{\Sigma}_\delta$.

If at least one of the correlations between the pairs $(\xi_j^S, \xi_j^{P(m)})$ or $(\delta_{ij}^S, \delta_{ij}^{P(m)})$ is not null, the selection mechanism is not ignorable, so unbiased estimation requires to fit both sets of equations simultaneously. It is worth to note

that in the multilevel case the selection mechanism can operate at different levels: (a) *subject level*: correlations between the pairs $(\delta_{ij}^S, \delta_{ij}^{P(m)})$; (b) *cluster level*: correlations between the pairs $(\xi_j^S, \xi_j^{P(m)})$. The signs of the correlations may be different at the two levels, giving rise to complex selection mechanisms. Moreover, ignoring the multilevel structure amounts to mix different aspects of the selection mechanism and might lead to wrong conclusions.

Note also that, whatever the selection mechanism, the random terms in the linear predictors of the multinomial logit model allow to relax the restrictive IIA (Independence of Irrelevant Alternatives) assumption (Skrondal and Rabe-Hesketh, 2003).

The parameters of the cluster level covariance matrix $\boldsymbol{\Sigma}_\xi$ are all identified, while for the parameters of subject level covariance matrix $\boldsymbol{\Sigma}_\delta$ the identification issue is more complex: the variance of δ_{ij}^S is obviously not identified, while the variances and covariances relative to the $\delta_{ij}^{P(m)}$, are in principle identified, but prone to empirical underidentification, unless some alternative specific covariate is included in the model (Skrondal and Rabe-Hesketh, 2003). Indeed, in the application $\boldsymbol{\Sigma}_\delta$ is found to be empirically not identified, so the δ_{ij}s are omitted.

3   Application

In the application the data set includes 2540 employed graduates and 56 degree programmes. The response of interest is the way of acquisition of the given skill: at university (reference category), at workplace or otherwise. The covariates used are the following. *Demographic*: gender, age at degree; *university career*: average mark of examinations (centered with respect to the mean of the degree programme), graduated with honors, duration index (ratio of time to graduate to legal duration); *job characteristics*: independent work, managerial post, public sector, temporary position, degree required for the job; *degree programme characteristics*: short degree.

Estimation is carried out by means of the `gllamm` procedure of Stata (Rabe-Hesketh *et al.*, 2001), which performs maximum likelihood estimation with adaptive Gaussian quadrature; the model selection is based on the likelihood ratio test.

For each of the eight skills included in the questionnaire, the polytomous model without selection is fitted. The skill with the highest estimated degree programme variance component is *Professional and technical abilities*, which is also the most interesting one for the University management. Therefore the joint model with selection is fitted only for this skill. Except for two students (about 0.1%), the non response to the acquisition question is always due to a negative response to the previous question on skill's use, so the source of bias resides only in the conditional nature of the acquisition question.

TABLE 1. Estimated probabilities from the acquisition model

Kind of graduate	$\mathrm{P}(Y_{ij}^P = m \mid \mathbf{x}_{ij}^P, \xi_j^{P(2)}, \xi_j^{P(3)})$		
and degree programme	University	Workplace	Otherwise
baseline	0.475	0.427	0.098
honors	0.575	0.347	0.078
self-employed work	0.534	0.354	0.113
managerial post	0.530	0.363	0.107
public sector	0.554	0.333	0.113
degree not required	0.284	0.556	0.159
short degree	0.573	0.379	0.049
high degree programme	0.269	0.529	0.201
low degree programme	0.681	0.280	0.039

The two cluster-level estimated correlations among the S and P sets of equations are jointly not significant (LRT=5.52, df=2, p-value=0.0633). This test may have a low power, however ignoring the selection mechanism causes only minor changes in the parameter estimates of the multinomial model. Therefore the analysis proceeds with the acquisition model alone, assuming an ignorable selection mechanism.

The cluster-level random parameters take the following values: $Var(\xi_j^{P(2)}) = 0.153$, $Var(\xi_j^{P(3)}) = 0.413$, $Corr(\xi_j^{P(2)}, \xi_j^{P(3)}) = 0.848$. Therefore, given the observed covariates, there is still much unexplained variability due to the degree programmes. Moreover the positive sign of the correlation implies that the second and third alternatives are jointly opposed to the first one. Table 1 reports the estimated probabilities for some combinations of the covariates: the *baseline* graduate is defined by setting all the covariates and random terms to zero; the row labelled *low (high) degree programme* corresponds to a graduate with all the covariates set to zero and each random term equal to minus (plus) twice the corresponding estimated standard error.

As for the covariates, the probability of acquisition during the degree programme is higher for graduates with honor and graduates with a short degree, while this probability significantly decreases if the degree is not required for the job. The job characteristics have little effect on the third alternative, while they substantially modify the probability of acquisition at the workplace.

4   Concluding remarks

In the application the hierarchical structure has a crucial role, while selection bias results negligible. However the outlined methodology can be

effectively used in situations where selection bias is an issue.

Currently we are carrying out some simulations to fully understand the implications of selection mechanisms that act in a hierarchical framework and to assess the power of the likelihood ratio test performed to evaluate the presence of selection.

Alternatively, selection bias can be treated following a sensitivity approach (Copas and Li, 1997), without relying on a single estimate for the parameters governing the selection mechanism. Bellio and Gori (2003) present an application of this approach in a multilevel setting.

The estimation algorithm based on adaptive numerical quadrature, used in the application, is accurate and flexible, but it requires long computational times, which increase rapidly with the model complexity. Many alternative estimation methods are possible, e.g. Bayesian MCMC and Maximum Simulated Likelihood (Train, 2003).

The analysis described in the paper is implicity conditional on the employment status of the graduates at the interview, so the results have to be referred only to the employed graduates. In order to evaluate the degree programmes with respects to the skills they give to all the graduates, it is necessary to take into account also the possible selection bias induced by the employment status.

References

Bellio, R. and Gori, E. (2003). Impact Evaluation of Job Training Programmes: Selection Bias in Multilevel Models, *Journal of Applied Statistics*, **30**, 893-907.

Borgoni, R. and Billari, F. C. (2002). A Multilevel Sample Selection Probit Model with an Application to Contraceptive Use. In: *Proceedings of the XLI meeting of the Italian Statistical Society*. Padova: CLEUP.

Copas, B. J. and Li, H. G. (1997). Inference for non-Random Samples (with discussion), *Journal of the Royal Statistical Society B*, **59**, 55-95.

Heckman, J. J. (1979). Sample selection bias as a specification error, *Econometrica*, **47**, 153-161.

Rabe-Hesketh, S., Pickles, A. and Skrondal, A. (2001). GLLAMM Manual. Technical Report 2001/01, Department of Biostatistics and Computing, Institute of Psychiatry, King's College, London.

Skrondal, A. and Rabe-Hesketh, S. (2003). Multilevel logistic regression for polytomous data and rankings, *Psychometrika*, **68**, 267-287.

Train, K. (2003). *Discrete Choice Methods with Simulation*. New York: Cambridge University Press.

Joint Modelling of Cluster Size and Binary and Continuous Outcomes

R. Gueorguieva[1]

[1] Division of Biostatistics, Department of Epidemiology and Public Health, Yale University School of Medicine, 60 College St P.O.Box 208034, New Haven, CT 06520-8034, USA

Abstract: In clustered designs often multiple outcome variables are collected for each individual. Some of the dependent variables may be measured at the individual level while others (for example cluster size) may be measured at the cluster level. It is both important and challenging to model all variables jointly taking into account the correlation between the variables. In this paper we consider a data example with a binary and continuous individual-level outcomes and an ordinal cluster-level variable, define a multivariate random effects model and obtain maximum likelihood estimates using standard software. We also compare bias in dose effect estimates when misspecifying the correlation structure of the random effects and when ignoring cluster size using a simulation study.

Keywords: maximum likelihood; multivariate response; random effects; repeated measures; Gaussian quadrature

1 Introduction

Joint modelling of multiple discrete and continuous outcomes presents challenges to investigators because of the need to model correlation between the outcomes within individual. The situation becomes even more complex when clustering is present and when both cluster-level and individual-level variables are present.

This paper is motivated by a developmental toxicity application (Price, Kimmel, Tyl and Marr, 1985). This was a study of the teratogenic effects of ethylene glycol conducted by the National Toxicology Program. During organogenesis pregnant mice were exposed to ethylene glycol at one of four different dose levels: 0, 0.75, 1.5 and 3 mg/kg. Fetal weight and a binary malformation indicator for each fetus within litter, and litter size were recorded. It was of interest to estimate the dose effect on adverse outcomes (malformation, low fetal weight). Descriptive statistics for the developmental toxicity data are available in Table 1.

A number of authors jointly analyzed the malformation and fetal weight outcomes. However only in the latest published Bayesian analysis (Dunson, Chen and Harry, 2003) and in the latest maximum-likelihood analysis

TABLE 1. Descriptive statistics for the developmental toxicity example.

Dose (g/kg)	Dams	Fetal Weight (g)		Malformation	
		Mean	SD	Number	Percent
0	25	0.972	0.098	1	0.34
0.75	24	0.877	0.104	26	9.42
1.50	22	0.764	0.107	89	38.86
3.00	23	0.704	0.124	126	57.08

(Gueorguieva, 2004) was litter size modelled as an additional dependent variable. As demonstrated by DCH ignoring litter size could lead to biased inferences although the extent of the bias might not be very large. In this paper we consider a model with separate litter level random effect for each outcome and a correlated probit formulation for litter size, and discuss how to obtain maximum likelihood estimates using the gllamm function in STATA or PROC NLMIXED in SAS. We also use a simulation study to compare bias in dose effect estimates when assuming a shared litter random effect instead of correlated random effects for the outcome variables and when ignoring litter size.

2   Model definition

Let y_{ij1} denote the weight of the jth fetus in litter i ($i = 1, ...I$, $j = 1, ...n_i$) and let $y_{ij2} = 1$ if the jth fetus in the ith litter is malformed and $y_{ij2} = 0$ otherwise. As usual we assume that there is a latent normal variable y_{ij2}^* underlying y_{ij2} such that $y_{ij2} = I(y_{ij2}^* > 0)$. Also, let s_i denote the size of litter i. Then the model we consider is defined as follows:

$$
\begin{aligned}
y_{ij1} &= \mu_1 + \alpha_1 x_i + \lambda_1 \xi_{i1} + \gamma_1 \eta_{ij} + \epsilon_{ij1} \\
y_{ij2}^* &= \mu_2 + \alpha_2 x_i + \lambda_2 \xi_{i2} + \gamma_2 \eta_{ij} + \epsilon_{ij2} \\
Pr(s_i \le k | x_i, \xi_{i3}) &= \Phi(\delta_k - \beta x_i - \lambda_3 \xi_{i3}),
\end{aligned}
$$

where $\boldsymbol{\xi} = (\xi_{i1}, \xi_{i2}, \xi_{i3})^T \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ is a vector of litter-specific random effects independent of the fetus-specific random effect η_{ij} and of the errors $\epsilon_{ij1} \sim N(0, \sigma_{e1}^2)$ and $\epsilon_{ij2} \sim N(0, \sigma_{e2}^2)$. For identifiability, the diagonal elements of $\boldsymbol{\Sigma}$ are assumed to be equal to 1, $\lambda_3 = 1$, and $\gamma_2 = \sigma_{e2} = \sqrt{0.5}$. The latter restriction means that the variance of the latent continuous variable underlying the malformation response is assumed to be one. The dose of ethylene glycol is denoted by x_i. The third equation above corresponds to a cumulative probit model for litter size with $k = 1, ...T - 1$ where T is the maximum litter size (16 in the data example). We require $\delta_1 < \delta_2 < ...\delta_{T-1}$. Note that correlations between fetal weight and litter size ($\rho(y_{ij1}, s_i)$), and between malformation and litter size ($\rho(y_{ij2}^*, s_i)$) arise from the correlated

random effects ξ_{i1}, ξ_{i2} and ξ_{i3}. Correlations between malformation and fetal weight measured on the same fetus within litter $(\rho(y_{ij1}, y_{ij2}^*))$ arise because of the random litter effects and of the common litter effect η_{ij}, while correlations between malformation measured on one fetus within a litter and weight of another fetus within a litter $(\rho(y_{ij1}, y_{ij'2}^*))$ are due only to the correlated random effects. This model is more general than the model Dunson et al. considered for this particular data example since they assumed a common litter effect for all outcomes ξ_i thus imposing a restrictive structure on the correlations. Both Dunson et al. and Gueorguieva used a continuation ratio formulation for cluster size to avoid having to place restrictions on the thresholds. However in the correlated probit model the thresholds can be reparametrized to avoid computational problems and the cumulative-probit formulation has the advantage of easier computation of correlations between litter size and fetal weight, and between litter size and malformation.

3   Maximum Likelihood Estimation

The model as defined above is a special case of the Generalized Linear Latent and Mixed Models (GLLAMM: Rabe-Hesketh, Skrondal and Pickles, 2001) and can be fitted using the gllamm function in Stata. Alternatively, the three-level model above can be rewritten as a two-level model by combining the fetus-level random effect γ_{ij} and the error ϵ_{ij1} for fetal weight, and by combining γ_{ij} and ϵ_{ij2} for malformation, thus creating a bivariate random error vector. The relationship between two-level and three-level formulations of models have been discussed by Grilli and Rampichini (2003) for ordinal data. The two-level formulation then allows the technique proposed by Gueorguieva (2004) to be used to fit this model in SAS using the *general* likelihood option in SAS PROC NLMIXED. Both gllamm in Stata and PROC NLMIXED in SAS obtain maximum-likelihood estimates using adaptive Gaussian quadrature.

4   Results

We compared the results from fitting the proposed model (Model 3) to the results from maximum likelihood estimation of the model with one shared random effect (Model 2: ξ_{i1}, ξ_{i2} and ξ_{i3} perfectly correlated) and to the results of the model considered previously by Gueorguieva and Agresti (2001) (Model 1: s_i dropped as a dependent variable). Dose of ethylene glycol was significantly associated both with decrease in fetal weight and with increase in the probability for malformation. The estimates of the parameters corresponding to the fetal weight variable were essentially the same regardless of which model was used. More pronounced differences were observed for the estimates corresponding to the malformation variable with estimates

TABLE 2. Maximum likelihood estimates for the parameters in the developmental toxicity example.

Parameter	Model 1 MLE(SE)	Model 2 MLE(SE)	Model 3 MLE(SE)
Weight			
Intercept (μ_1)	0.944(0.015)	0.944(0.015)	0.945(0.015)
Dose (α_1)	-0.087(0.009)	-0.087(0.009)	-0.087(0.009)
Factor loading (λ_1)	0.088(0.007)	0.088(0.007)	0.089(0.007)
Error SD ($\sigma_{\epsilon 1}$)	0.095(0.002)	0.095(0.002)	0.094(0.002)
Malformation			
Intercept (μ_2)	-2.331(0.201)	-2.085(0.146)	-2.307(0.198)
Dose (α_2)	0.917(0.103)	0.804(0.076)	0.915(0.102)
Factor loading (λ_2)	-0.788(0.007)	-0.561(0.075)	-0.779(0.098)
Litter size			
Dose (β)	–	-0.286(0.100)	-0.384(0.136)
Factor loading (λ_3)	–	-0.277(0.119)	1.00(0.00)

based on the model with the simpler random effects structure being significantly smaller. These results are consistent with results obtained using a continuation ratio formulation for litter size. To investigate the extent of the bias due to misspecifying the random effects structure and the bias due to ignoring cluster size we performed a small simulation study.

5   Simulation study

We simulated 500 data sets according to the most general model (Model 3) and we set parameters to be equal to the MLEs from Model 3 in the data example. We fitted all three models defined above to each data set. Table 3 contains bias and average SE estimates for the regression parameters according to the three models considered in the simulation study. Our results confirm the observation that in this particular application the bias in dose effect estimates for the binary response is significantly larger when considering an overly simplified correlation structure than when omitting litter size as a dependent variable.

6   Discussion

This paper demonstrates how to obtain maximum-likelihood estimates in a repeated measures example with individual-level binary and continuous variables and cluster size as another dependent variable. A correlated-probit model formulation for cluster size is both computationally and interpretationally convenient. It is easy to extend the suggested model to other

TABLE 3. Bias and average standard error for the parameters in the simulation study.

Parameter	Model 1 Bias(MSE)	Model 2 Bias(MSE)	Model 3 Bias(MSE)
Continuous			
μ_1	-0.0004(0.016)	-0.005(0.016)	0.0002(0.016)
α_1	-0.0004(0.009)	0.003(0.009)	-0.001(0.009)
λ_1	-0.002(0.008)	-0.003(0.007)	-0.002(0.008)
Binary			
μ_2	-0.025(0.199)	0.307(0.112)	-0.019(0.199)
α_2	-0.007(0.099)	-0.067(0.056)	0.005(0.099)
λ_2	0.016(0.106)	0.345(0.055)	0.016(0.105)

mixtures of binary, ordinal and continuous dependent variables either at individual or at cluster level. Our simulation study underline the importance of careful selection of the random effects structure for inferences on the regression parameters.

References

Dunson, D., and Chen B., and Harry J. (2003). A Bayesian Approach for Joint Modeling of Cluster Size and Subunit-Specific Outcomes. *Biometrics*. **59**, 521-530.

Grilli, L., and Rampichini, C. (2003). Alternative specifications of multivariate multilevel probit ordinal response models. *Journal of Educational and Behavioral Statistics*. **28**, 31-44.

Gueorguieva, R.V. (2004). Comments about Joint Modeling of Cluster Size and Binary and Continuous Subunit-Specific Outcomes. In review for Biometrics

Gueorguieva, R.V., and Agresti, A. (2001). A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association*. **96**, 1102-1112.

Price, C.J., and Kimmel, C.A., and Tyl, R.W. and Marr, M.C. (1985). The Developmental Toxicity of Ethylene Glycol in Rats and Mice. *Toxicological Applications in Pharmacology*. **81**, 825-839.

Rabe-Hesketh, S., and Pickles A., and Skrondal A. (2001). GLLAMM: A class of models and a Stata program. *Multilevel Modelling Newsletter*. **13**, 17-23.

Overdispersion in Wadley's Problem

Linda M. Haines[1] and Kerry Leask[1]

[1] School of Mathematics, Statistics and Information Systems, University of Kwa-Zulu-Natal, Private Bag X01, Scottsville 3209, Pietermaritzburg, South Africa

Abstract: Wadley's problem relates to dose-response experiments in which the number of individuals surviving a given dose is recorded but the number originally present in the system is unknown. The situation can be modelled by assuming that the number of individuals initially present is Poisson and that the number of individuals surviving, given the number originally present, is binomial. It then follows that the number of individuals surviving is Poisson with parameter proportional to the probability of survival. In the present study an approach to the modelling of overdispersion in Wadley's problem based on the assumption that the probability of survival is beta distributed is introduced and follows closely the development of the beta-binomial paradigm. The resultant beta-Poisson distribution is reviewed and estimation of the model parameters within the dose-response context is illustrated by means of data drawn from a study on anti-malarial drugs.

Keywords: Wadley's problem; Overdispersion; Beta-Poisson distribution; Maximum Likelihood; Malaria data.

1 Introduction

Wadley (1949) first considered modelling dose-mortality data for which the number of organisms initially exposed to a treatment is unknown and must therefore be estimated from a control sample. This phenomenon frequently emerges in dose-response experiments and is aptly termed Wadley's problem. Wadley (1949) assumed that the number of organisms treated follows a Poisson distribution, while Anscombe (1949) introduced the notion of using the negative binomial distribution rather than the Poisson as a means of accommodating overdispersion. More recently Baker, Pierce and Pierce (1980) and Smith and Morgan (1989) developed GLIM and GENSTAT macros for use in analyzing overdispersed Wadley-type data and their work was consolidated in the paper by Morgan and Smith (1992). In the present study a new approach to accommodating overdispersion in Wadley's problem based on the beta-Poisson distribution is introduced and is illustrated by means of data taken from an antimalarial drug study.

TABLE 1. Data for malaria parasites exposed to the antimalarial drug Halofantrine.

| Drug conc. | PARASITAEMIA | | | |
(μ/l)	Count 1	Count 2	Count 3	Mean
0	4957	5065	5010	5011
1	5193	4897	4816	4969
2	4590	4516	4223	4443
4	3615	3356	3102	3357
8	914	816	657	796
16	49	12	12	18
32	23	30	19	24
64	33	88	62	61

2 Malaria data

Blood samples infected with *Plasmodium Falciparum* were taken from a Gambian malaria sufferer between July 1984 and February 1987. The samples were treated with varying concentrations of the antimalarial drug, Halofantrine, and the number of parasites surviving was recorded. Three batches were exposed to each dose of the drug and the results are summarized in Table 1. The data were collected by researchers from the Medical Research Council in Durban, South Africa, involved in the Malaria National Program and are extracted from the Masters thesis of Gouws (1995, p.98).

3 Preliminaries

Let $y_{cj}, j = 1, \ldots, n_c$, denote an observation from a control group in which the drug is not administered and suppose that the number of parasites in such a group follows a Poisson distribution with parameter τ. Let y_{ij} refer to the number of surviving parasites at a non-zero concentration d_i of the drug, $i = 1, \ldots, D$ and $j = 1, \ldots, n_i$. For each dose d_i, the log-dose is given by $x_i = \log d_i$ and the associated probability of death of a parasite is denoted by $p_i, i = 1, \ldots, D$. Wadley (1949) showed that if the number of organisms treated at log-dose x_i is assumed to follow a Poisson distribution with parameter τ then the number of organisms surviving will also follow a Poisson with the parameter $\tau(1 - p_i), i = 1, \ldots, D$. Furthermore the probability of death can be modelled using the logit function $\ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta x_i$ where α and β are unknown parameters, so that the expected number of parasites surviving for a log-dose x_i is $\frac{\tau}{1 + e^{\alpha + \beta x_i}}, i = 1, \ldots, D$. The overall formulation is therefore that of a generalized nonlinear model.

This model was fitted to the malaria data in Table 1 using the method of maximum likelihood and the resultant parameter estimates were obtained as $\hat{\tau} = 5011.22, \hat{\alpha} = -4.523$ and $\hat{\beta} = 6.748$. The deviance as compared with the maximal model, where $Y_{ij} \sim \text{Poisson}(\lambda_{ij})$, $i = 1, \ldots, D$ and $j = 1, \ldots, n_i$, was found to be 1544.784. This value is highly significant based on a χ^2 distribution with 21 degrees of freedom and indicates that the model provides a poor fit to the data. An examination of the residuals further showed that the apparent lack of fit is due to the presence of overdispersion in the data. In order to accommodate such overdispersion, Anscombe (1949) mirrored Wadley's findings for the Poisson model with results based on the negative binomial distribution. Anscombe's model, with the probability of death described by a logit function, was therefore fitted to the malaria data using maximum likelihood. The resultant deviance was however found to be very highly significant and the residual plots again indicated the presence of overdispersion. It should be noted that replication and batch effects in the blood samples could well contribute to the observed overdispersion in the data. Gouws (1995) examined this issue particularly carefully and concluded that, on the basis of the experimental procedures followed, the prescence of such effects could not be justified.

4   Beta-Poisson model

Suppose that a random variable Y follows a Poisson distribution with parameter $\tau(1-p)$ and that the parameter p in turn follows a beta distribution with parameters a and b where $a > 0$ and $b > 0$. Then Y is said to follow a beta-Poisson distribution with probability density function given by

$$P(Y = y) = \frac{\tau e^{-\tau}}{y!} \frac{\Gamma(a + b)\Gamma(b + y)}{\Gamma(a + b + y)\Gamma(b)} \, {}_1F_1(a, a + b + y; \tau)$$

where ${}_1F_1(\ )$ represents the confluent hypergeometric or Kummer function. This distribution is a variant of the Poisson-beta distribution introduced by Bhattacharya and Holla (1965) and described, with further details and references, in Johnson, Kotz and Kemp (1992). The beta-Poisson distribution can be used within the context of Wadley's problem by following the classical approach to the beta-binomial model described in Morgan (1992, Section 6.3). Specifically, at log-dose x_i, with probability of death p_i following a beta distribution with parameters a_i and b_i, a logit function can be used to model the expected value of p_i as $\pi_i = \dfrac{a_i}{a_i + b_i} = \dfrac{e^{\alpha + \beta x_i}}{1 + e^{\alpha + \beta x_i}}$ and an additional shape parameter $\theta = \dfrac{1}{a_i + b_i}$ can be introduced for $i = 1, \ldots, D$. Then the log-likelihood for the model-data setting is given by

$$
\begin{aligned}
l \;=\; & \sum_{j=1}^{n_c} \ln \left\{ \frac{e^{-\tau} \tau^{y_{cj}}}{y_{cj}!} \right\} + \\
& \sum_{i=1}^{D} \sum_{j=1}^{n_i} \ln \left\{ \sum_{s=0}^{\infty} \frac{e^{-\tau} \tau^{y_{ij}}}{y_{ij}!} \frac{\Gamma(a_i + b_i)\Gamma(b_i + y_{ij})\Gamma(a_i + s)}{\Gamma(a_i + s + b_i + y_{ij})\Gamma(a_i)\Gamma(b_i)} \frac{\tau^s}{s!} \right\}
\end{aligned}
$$

where $a_i = \dfrac{\pi_i}{\theta}$ and $b_i = \dfrac{1 - \pi_i}{\theta}$ and the Kummer function is expressed as an infinite sum. Maximum likelihood estimates of the parameters were obtained for the malaria data by maximizing an approximation to the function l obtained by appropriately truncating the infinite sum and were given by $\hat{\tau} = 5035.73, \hat{\theta} = 0.012, \hat{\alpha} = -4.047$ and $\hat{\beta} = 6.779$. The deviance as compared with the maximal model described earlier was found to be 114.378 with a P-value very close to zero and is thus highly significant. The beta-Poisson model does not therefore provide an entirely satisfactory fit to the malaria data. However this observed deviance does indicate that the beta-Poisson model is a vast improvement on the Poisson and negative binomial models described in Section 3 in that it reduces the deviance of the former model by 1430.406 at the expense of just 1 degree of freedom and of the latter by 1373.742 with no change in the degrees of freedom.

5   Conclusions

A new approach to modelling overdispersion in Wadley's problem which is based on the beta-Poisson distribution is introduced. The method is broadly appealing and builds on the framework of the well-known beta-binomial model. There is much scope for further work. Thus it is of some interest to describe fully the properties of the beta-Poisson distribution and of the associated maximum likelihood estimates. In a broader context it is possible to extend some of the ideas for accommodating overdispersion in binomial models to Wadley's problem setting, as for example the approach based on random coefficients described in Aitkin (1996) and the models discussed in Lindsey and Altham (1998).

References

Aitkin M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, **6**, 251-262.

Anscombe F.J. (1949). The statistical analysis of insect counts based on the negative binomial distribution. *Biometrics*, **5**, 165-173.

Baker R.J., Pierce C.B., and Pierce J.M. (1980) Wadley's problem with controls. *GLIM Newsletter*, **2**, 29-30.

Bhattacharya S.K., and Holla M.S. (1965). On a discrete distribution with special reference to the theory of accident proneness. *Journal of the American Statistical Association*, **60**, 1060-1066.

Gouws E. (1985). *Drug Resistance in Malaria Research : the Statistical Approach*. M.Sc. thesis, University of Natal, South Africa.

Johnson N.L., Kotz S., and Kemp A.W. (1992). *Univariate Discrete Distributions*. 2nd Edition. New York: Wiley.

Lindsey J.K., and Altham P.M.E. (1998). Analysis of the human sex ratio by using overdispersion models. *Applied Statistics*, **47**, 149-157.

Morgan B.J.T., and Smith D.M. (1992). On Wadley's problem with overdispersion. *Applied Statistics*, **41**, 349-354.

Smith D.M., and Morgan B.J.T. (1990). Extended models for Wadley's problem. *GLIM Newsletter*, **18**, 21-35.

Wadley F.M. (1949). Dosage-mortality correlation with number treated estimated from a parallel sample. *Annals of Applied Biology*, **36**, 196-202.

Model Selection for P-spline smoothing using Akaike Information Criteria

Carrie Wager[1], Florin Vaida[1] and Göran Kauermann[2]

[1] Department of Biostatistics, Harvard School of Public Health, Boston USA;
cwager@hsph.harvard.edu and vaida@sdac.harvard.edu
[2] Department of Economics and Business Administration, University Bielefeld,
Bielefeld, Germany; gkauermann@wiwi.uni-bielefeld.de

Abstract: Penalized regression splines can be conveniently fit using software
and theory borrowed from linear mixed effects models. This has led to a boom
in the practical application of complex models having multiple and/or hierar-
chical smooth terms. We consider selecting the composition of smooth terms in
additive models by using two alternative formulations of the Akaike Information
Criterion (AIC) that are based on the marginal versus conditional likelihood. The
marginal likelihood provides the conventional inference for linear mixed effects
models, whereas a conditional perspective is traditionally used for choosing the
optimal smoothing parameter. Through simulation we find that in moderately
large samples, both the conditional and marginal formulations of AIC perform
extremely well at detecting the function which generated the data. The marginal
AIC does better for simple functions and in small samples, whereas the condi-
tional AIC does better at detecting a true function which has a complex hierar-
chical formulation. We provide examples of two real applications which motivate
this collaborative work: the first compares a penalized spline to the standard
parametric nonlinear pharmacokinetics model to assess the adequacy of its fit,
and the second involves selecting the level at which spatial intensity should be
modeled in a hierarchical ANOVA model of neuronal activation patterns in phar-
macological brain imaging.

Keywords: Penalized Spline; Model Selection; Conditional versus Marginal In-
ference; Variance Component Selection.

1 Introduction

Assume that we have data arising from a simple smoothing model:

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where y_i is the response for the ith subject, x_i is a measured scalar co-
variate, $f(x_i)$ is a smooth function of x_i, and $\epsilon_1, \ldots, \epsilon_n$ are error terms
with mean zero and variance σ^2. Using the mixed-model formulation of
penalized spline smoothing, f can be modeled using a linear combination

of covariates and parameters:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \sum_{k=1}^{K} z_k(x) u_k, \qquad (2)$$

where $(\beta_0, \beta_1, \beta_2)^T$ is a vector of fixed effects, $z_1(x), \ldots, z_K(x)$ are smooth basis terms that model the curvature in $f(x)$, and $u_1, \ldots, u_K \sim \mathcal{N}(0, \lambda\sigma^2)$ are independent random effects where the parameter $\lambda = \mathrm{var}(u_k)/\sigma^2$ controls the amount of smoothing.

Because they have a dimension K of the smoothing basis which is typically much smaller than the number of obervations n, penalized regression splines can be viewed as 'low rank' approximations to smoothing splines (Wahba 1990). Such models were made popular by Eilers & Marx (1996) who coined the term 'P-spline' when they introduced penalties to the popular B-spline techniques used in regression. From a different angle, Hastie (1996) developed 'pseudosplines' which reduced the rank of traditional smoothing splines by truncating an eigendecomposition of the smoothing basis. Our concept of P-splines encompasses all of these variations: regardless of the form of the smoothing basis $\{z_k(x)\}$, we fit model (2) for (1) using the machinery of a linear mixed effects model (Ruppert et. al. 2003).

Recent applied work draws on the convenience of the linear mixed effects model framework to extend P-splines to additive models with interaction between design factors and the smooth terms (Brumback & Rice 1998, Coull et. al, 2001, Kammann & Wand 2003, Wager et. al. 2004). In these cases, model (2) for (1) may have multiple and/or hierarchically-nested smooth terms, such as in the model

$$y_{i\ell} = f_1(x_{i\ell}) + f_{1\ell}(x_{i\ell}) + f_2(w_{i\ell}) + \ldots + \epsilon_{i\ell} \qquad (3)$$

where x and w are distinct covariates and we denote $\ell = 1, \ldots, L$ as group levels of a design factor where f_ℓ is a smooth level-specific deviation from the mean curve f. In (3), the main functions f_1 and f_2 have distinct smoothing parameters λ_1 and λ_2, whereas the set of group-specific functions $f_{11}, \ldots, f_{1L}$ typically share a common smoothing parameter λ_{11} over all groups.

2   Model selection

Given several competing models comprised of different subsets of smooth terms, our goal is to choose a model which provides the best predictive accuracy for future data arising from the true distribution. Perhaps $f_1(x)$ is highly correlated with $f_2(w)$ in (3), and we need to choose one function that provides a better model. Additionally, we may consider models which smooth at different levels of a design hierarchy, where model (3) is compared with both a common-curve model (1), and a model that replaces $f_{1\ell}$

in (3) with a constant factor α_ℓ. These types of model comparisons are a bit more complex than simple covariate selection because the competing models differ in both the composition of the regression parameters (β, α, u) which affect the conditional mean of the response, as well as the composition of the smoothing parameters $(\lambda_1, \lambda_{11}, \lambda_2)$ which affect its marginal variance.

3   Akaike information criteria

The goal of Akaike's (1973) information criterion is to minimize the the expected 'distance' between the true density function and the best model for a given set of data. This happens to be equivalent to maximizing the predictive likelihood $T = E_y E_{y^\star}(\ell(\hat{\theta}(y)|y^\star))$ where $y^\star$ is a new observation from the true distribution of y. While this does criterion does not fundamentally require that the true distribution of y necessarily be in the class of models for which the log-likelihood $\ell(\theta|y)$ is being maximized, typically it is assumed that the truth is in the class of models being fit in order to facilitate estimation. A general formula this criterion is:

$$\text{AIC} = -2\ell(\hat{\theta}|y) + 2 \cdot \text{bias} \tag{4}$$

where the bias term results from using the expected maximized likelihood $E_y(\ell(\hat{\theta}|y))$ to estimate the maximized predictive likelihood T. For a P-spline smoothing model which is fit using linear mixed model machinery, the criterion (5) can be formulated using either the conditional likelihood, where the parameters u are considered to be known:

$$\text{c}\ell(\beta, \sigma^2|y, u)) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{||y - X\beta + Zu||^2}{2\sigma^2}$$

or the marginal likelihood which averages over the distribution of the u's:

$$\text{m}\ell(\beta, \sigma^2|y) = \log E_u(\exp(\text{c}\ell(\beta, \sigma^2|y, u)).$$

This leads to the two alternative formulations of AIC:

$$\begin{aligned}
\text{mAIC} &= -2\log(\text{m}\ell(\hat{\beta}, \hat{\sigma}^2|y) + 2p \\
\text{cAIC} &= -2\log(\text{c}\ell(\hat{\beta}, \hat{\sigma}^2|y, \hat{u}) + 2\rho(\theta)
\end{aligned}$$

where, heuristically, the bias term p in the mAIC turns out to be the number of unknown parameters in the marginal likelihood, and the bias term $\rho(\theta)$ in the cAIC is the 'effective' number of unknown parameters in the conditional likelihood, and can be easily computed by taking the trace of the smoothing matrix.

4   Simulations

To compare the overall properties of mAIC versus cAIC for model selection, we consider two modelling scenarios: The first scenario compares models that have correlated smoothing terms, which we generate based on true data $y_i \sim \mathcal{N}(f_1(x_i), \sigma^2)$ where x_i is highly correlated with another covariate, w_i. We then fit the competing models

$$\begin{aligned} \mathcal{M}_A : & \quad y_i = f_1(x_i) + \epsilon_i \\ \mathcal{M}_B : & \quad y_i = f_2(w_i) + \epsilon_i \end{aligned}$$

and repeat this simulation for several levels of correlation between x and w, a range of small to large sample sizes, a range of residual errors, and several true nonlinear mean curves that have varying complexity.

The second scenario considers hierarchically-nested smooth terms, where we generate three alternative true models for the data:

$$\begin{aligned} \mathcal{M}_C : & \quad y_{i\ell} = f(x_{i\ell}) + \epsilon_{i\ell} & \text{(common curve)} \\ \mathcal{M}_D : & \quad y_{i\ell} = \alpha_\ell + f(x_{i\ell}) + \epsilon_{i\ell} & \text{(subject-specific intercepts)} \\ \mathcal{M}_E : & \quad y_{i\ell} = f(x_{i\ell}) + f_\ell(x_{i\ell}) + \epsilon_{i\ell} & \text{(subject-specific curves).} \end{aligned}$$

We repeat this simulation for permutations of wide and narrow distances between the group-specific curves, a range of residual errors, and true nonlinear mean curves having varying complexity. In each iteration, we fit each of the three models corresponding to $\mathcal{M}_A$, $\mathcal{M}_B$, and $\mathcal{M}_C$ to each of these three truths.

Overall, we find that in moderately large samples, both the conditional and marginal formulations of AIC perform equally well at detecting the function which generated the data. The smoothing parameter chosen by mAIC (equivalent to marginal maximum likelihood) tends to result in a smoother fit than the smoothing parameter chosen by cAIC, lending further support to theoretical results previously reported in Kauermann (2004). The mAIC performs better than cAIC for simple functions and in small samples, whereas the cAIC does better at detecting a true function which has a complex hierarchical form.

5   Examples

We provide examples based on two real-data applications which motivate this collaborative work. The first example, motivated by Vaida & Blanchard (2004) compares a penalized spline to the standard nonlinear parametric pharmacokinetics model to assess the adequacy of its fit. The second example, motivated by Wager et. al. (2004) involves selecting the level at which spatial intensity should be modeled in a hierarchical ANOVA of replicated patterns of neuronal activation in brain imaging.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: *Breakthroughs in Statistics*. 610-624, Springer-Verlag.

Brumback, B. A. and Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, **93**, 961-976.

Coull, B. A., Ruppert, D. and Wand, M. P. (2001). Simple incorporation of interactions into additve models. *Biometrics*, **57**, 539-545.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89-102.

Hastie, T. (1996). Pseudosplines. *Journal of the Royal Statistical Society, Series B* **58**, 379-396.

Hurvich, C. M., Simonoff, J. S. and Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion, *Journal of the Royal Statistical Society, Series B*, **60**, 271-293.

Kammann, E.E. and Wand, M.P. (2003). Geoadditive models. *Applied Statistics*, **1**, 1-18.

Kauermann, G. (2004). A note on smoothing parameter selection for penalized spline smoothing. *Journal of statistical planning and inference*, in press.

Ruppert, D., Wand, M. and Carroll, R. J. (2003). *Semiparametric Regression*, Wiley.

Simonoff, J. S. and Tsai, C.-L. (1999). Semiparametric and additive model selection using an improved Akaike information criterion. *Journal of Computational and Graphical Statistics*, **8**, 22-40.

Vaida, F. and Blanchard, S. (2004). Conditional Akaike information for mixed effects models. (Submitted for publication).

Wager, C. G., Coull, B. A. and Lange, N. (2004). Modelling spatial intensity for replicated inhomogeneous point patterns in brain imaging. *Journal of the Royal Statistical Society, Series B*, **66**, 1-18.

Wahba, G. (1990) *Spline Functions for Observational Data*. SIAM.

A Bayesian Accelerated Failure Time Model with a Normal Mixture as an Error Distribution

Arnošt Komárek[1] and Emmanuel Lesaffre[1]

[1] Catholic University Leuven, Biostatistical Centre, Kapucijnenvoer 35, B–3000, Leuven, Belgium

Abstract: A Bayesian version of the accelerated failure time model for possibly dependent data is proposed. The error distribution is modelled via a normal mixture with unknown number of components, in practice avoiding any distributional assumptions concerning the event times. The approach is illustrated on a CGD data.

Keywords: Censored Data; Clustered Data; Regression; Reversible Jump Markov Chain Monte Carlo.

1 Introduction

In the survival analysis, the accelerated failure time model (AFT) is a worthwhile alternative to the Cox's relative risks (RR) model. It was further suggested by Keiding et al. (1997) that including a random effect in the AFT model for clustered data would be an interesting alternative to the frailty RR model.

The AFT model with a random effect specifies that the effect of a vector of fixed covariates $\mathbf{x}_{il}$ together with a random effect b_i act additively on the logarithm of the time to event T_{il} of the lth observational unit in the ith cluster as

$$\log(T_{il}) = Y_{il} = b_i + \beta^T \mathbf{x}_{il} + \varepsilon_{il}, \quad i = 1, \ldots, N, \ l = 1, \ldots, n_i, \quad (1)$$

where ε_{il} is the error term with a density $f(e)$ and β is a vector of regression parameters. Unlike the area of uncensored data where the normal distribution is the most used error distribution, non- or semi-parametric procedures are generally preferred in the survival analysis.

Richardson and Green (1997) suggested to represent a non-standard density as a mixture of normals with the number of mixture components as well as all mixture parameters (weights, means and variances) being treated as unknown quantities in a Bayesian manner. We adapted their method to represent a density of the error term in the regression model with censored observations (AFT model). Thus, our model is, in fact, completely

parametric. However, due to the well known fact that under mild conditions, a continuous density can be approximated as precisely as desired by a normal mixture, in practice, we do not make any distributional assumptions regarding the error term. The advantage of our approach (at least in some situations) compared to completely non-parametric techniques is the fact that it produces an estimate of the error density which can be easily understood and compared (via plots) to standard parametric densities.

2   Bayesian model and Inference

To put several types of censoring (right, left and interval) into one framework we will assume that the observed log-event time of the (i, l)th unit is given by a pair (y_{il}^L, y_{il}^U), $-\infty \leq y_{il}^L \leq y_{il}^U \leq \infty$. For an uncensored observation, $y_{il}^L = y_{il}^U$, for a right censored observation, $y_{il}^U = \infty$ and for a left censored observation, $y_{il}^L = -\infty$. Further, let y_{il} denote (in the case of censoring unknown) value of the log-event time of the (i, l)th unit in the data set.

The density $f(e)$ of the error term ε_{il} in the model (1) is specified as $f(e) = \sum_{j=1}^k w_j \varphi(e | \mu_j, \sigma_j^2)$, with $\varphi(\cdot | \mu_j, \sigma_j^2)$ being a density of a normal distribution with mean μ_j and variance σ_j^2. Note that the number of mixture components, k, is unknown as well as mixture weights $w = (w_1, \ldots, w_k)^T$, means $\mu = (\mu_1, \ldots, \mu_k)^T$ and variances $\sigma^2 = (\sigma_1^2, \ldots, \sigma_k^2)^T$. To describe the model, we will, latently, assume that each conditional (given β, $\mathbf{x}_{il}$ and b_i) residual $e_{il} = y_{il} - b_i - \beta^T \mathbf{x}_{il}$ is distributed according to one mixture component. Let r_{il} be an index of this component. Since the density $f(e)$ is not necessarily of zero mean we do not allow an inclusion of the intercept term in the covariate vector $\mathbf{x}_{il}$.

The Bayesian model we use has a clear hierarchical structure and it is best described by a direct acyclic graph (DAG) where the squared boxes represent observed quantities or fixed hyperparameters and circles the unknowns. The DAG for our model is shown on Figure 1. Finally, we point out that although the censoring appears in the model there is no need to model it explicitly provided the censoring is independent. In that case, only its observed realization is needed to get a posterior distribution of quantities of interest.

We use the following prior assumptions determining the model given by DAG on Figure 1. Poisson distribution with mean λ truncated at k_{max} is assumed for number of mixture components k. Symmetric k-dimensional Dirichlet distribution with all 'prior sample sizes' equal to a hyperparameter δ is adopted for mixture weights w. It is further assumed that mixture means μ_j and variances σ_j^2 are all drawn independently, with normal $N(\xi, \kappa)$ priors for μ_j's and inverse-gamma $IG(\zeta, \eta)$ priors for σ_j^2's. Since the whole model is invariant to permutations of the labels $j = 1, \ldots, k$, we restrict
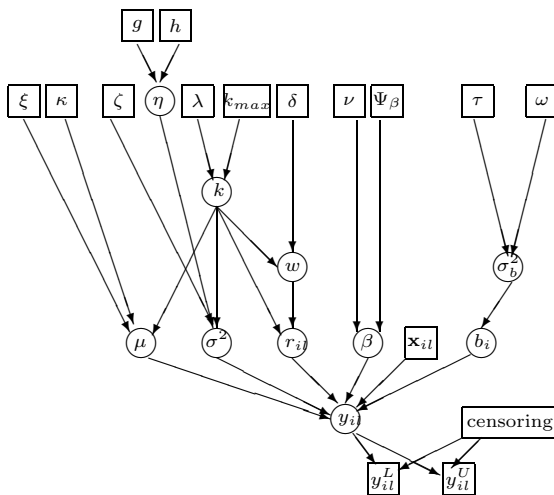
FIGURE 1. DAG for the Bayesian AFT model.

the joint prior distribution of a vector μ to the set $\{\mu : \mu_1 < \cdots < \mu_k\}$ for identifiability.

In the mixture context, it is not possible to be fully non-informative and to obtain proper posterior distributions. However, weak priors for the mixture parameters in our regression setting can be obtained in the following way. First, we fit an AFT model with a normal error distribution, e.g., using standard maximum-likelihood techniques. The hyperparameter ξ is then set to an estimated intercept value. The hyperparameter κ is set to a multiple of R^2 where R denotes an estimated scale from the maximum-likelihood fit. Since the knowledge of R does not imply much about the size of each single σ_j^2, an additional level of hierarchy by allowing η to follow a gamma distribution $G(g, h)$ with $\zeta > 1 > g$ and h being a small multiple of $1/R^2$ was suggested by Richardson and Green (1997) to express the belief that the σ_j^2's are similar, without being informative about their absolute size.

From the definition of latent allocation variables r_{il}, their prior distribution is given by $P(r_{il} = j \mid k, w) = w_j$.

The prior assumptions for the regression part of the model used in this paper are rather standard in the area of a hierarchical modelling. All components of the vector $\beta = (\beta_1, \ldots, \beta_p)^T$ are a priori independent, each with normal distribution $N(\nu_m, \psi_m)$. The matrix Ψ_β from the DAG is thus a diagonal matrix with $\psi_1, \ldots, \psi_m$ on the diagonal. The random effects b_i

are assumed a priori to be i.i.d. across clusters, with normal distribution $N(0, \sigma_b^2)$. The variance σ_b^2 of random effects has a priori an inverse-gamma distribution $IG(\tau, \omega)$ where τ and ω are fixed hyperparameters, typically chosen such that the ratio τ/ω^2 is high.

The list of conditional distributions from the DAG continues by an explicit specification of a distribution of (unobserved) log-event times y_{il} given μ, σ^2, r_{il}, β, $\mathbf{x}_{il}$ and b_i. This is a product of independent normal distributions with mean $\mu_{r_{il}} + \beta^T \mathbf{x}_{il} + b_i$ and variance $\sigma_{r_{il}}^2$ for (i,l)th observation. Finally, the conditional joint density of limits of observed intervals (y_{il}^L, y_{il}^U) given censoring and latent true data is given by the expression $p(y_{il}^L, y_{il}^U \mid y_{il}, \text{censoring}) \propto I[y_{il}^L < y_{il} \leq y_{il}^U] \cdot p(y_{il}^L, y_{il}^U \mid \text{censoring})$. Note that $p(y_{il}^L, y_{il}^U \mid \text{censoring})$ does not have to be specified explicitly to draw an inference based on posterior distribution, i.e. on the distribution

$$
p\Big(\{y_{il}\}, w, \mu, \sigma^2, \{r_{il}\}, k, \eta, \beta, \{b_i\}, \sigma_b^2 \;\Big|
$$
$$
\{(y_{il}^L,\ y_{il}^U)\}, \text{censoring}, \{\mathbf{x}_{il}\}, \xi, \kappa, \zeta, g, h, \lambda, k_{max}, \delta, \nu, \Sigma_\beta, \tau, \omega\Big).
$$

The inference in a Bayesian modelling is based on the quantities derived from above posterior distribution (posterior means, quantiles etc.). To get the posterior quantities of an interest, a Markov chain Monte Carlo technique is exploited here. The details of the sampling algorithm related to the update of the mixture parameters can be found in Richardson and Green (1997). The remaining quantities related to the regression model are sampled using a Gibbs move.

The sampling algorithm as well as some tools for computing the posterior quantities were implemented as a set of R functions with time consuming parts being performed by a C++ compiled code. These routines are available upon request from the first author.

3   Illustration: CGD Data

We illustrate our approach on the analysis of the data set from a placebo-controlled randomized trial of gamma inferon in patients with chronic granulotomous disease (CGD). The data set can be found in Appendix D.2 of Fleming and Harrington (1991). There were 128 patients randomized to either gamma inferon ($n = 63$) or placebo ($n = 65$). The data for each patient gives the time from study entry to initial and any recurrent serious infections. There is a minimum of one record per patient, with a total of 203 records. The data set has been analysed by various authors, including Vaida and Xu (2000) who used the relative risks model with a normal random effect for a patient on log–hazard scale.

We fitted the AFT model (1) with time from entry or previous infection to the next infection as a response, random effect term for a patient and covariates significant from an oridinary Cox regression as reported by Vaida

TABLE 1. Estimates from the CGD data.

Parameter	Poster. mean	95% cred. int.
treatment (yes)	1.275	$(0.486, 2.167)$
inherit (autosomal recessive)	-0.912	$(-1.806, -0.047)$
age (years)	0.046	$(0.006, 0.090)$
corticosteroids (yes)	-2.617	$(-5.260, -0.246)$
prophylactic antibiotics (yes)	1.072	$(0.071, 2.174)$
gender (female)	1.406	$(0.120, 2.823)$
hosp1 (US – other)	0.367	$(-0.532, 1.319)$
hosp2 (Europe – Amsterdam)	1.547	$(0.135, 3.114)$
hosp3 (Europe – other)	1.145	$(-0.077, 2.486)$
Mean of the error density	3.963	$(2.382, 5.624)$
Scale of the error density	2.007	$(1.291, 3.745)$
σ_b	0.626	$(0.043, 1.355)$

and Xu (2000). Vague priors were used for all parameters. Posterior means
and 95% posterior credibility intervals for regression parameters, mean and
scale of the error distribution and a standard deviation of the random effect
are found in Table 1.

The results we obtained consent qualitatively with the results of the Cox
model with the random effects of Vaida and Xu (2000). Further, as well as
these authors we observe that the random effects of patients with different
numbers of total infections are quite different suggesting that patients with
more infections are different from patients with less infections and that this
difference cannot be explained by covariates included in the model.

References

Fleming, T. R., and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* New York: Wiley.

Keiding, N., Andersen, P. K., and Klein, J. P. (1997). The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in Medicine,* **16,** 215–225.

Richardson, S., and Green, P. J. (1997). On Bayesian analysis of mixtures with unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B,* **59,** 731–792.

Vaida, F., and Xu, R. (2000). Proportional hazards model with random effects. *Statistics in Medicine,* **19,** 3309–3324.

The misclassification SIMEX

Helmut Küchenhoff[1], Emmanuel Lesaffre[2] and Samuel
M. Mwalili[2]

[1] Department of Statistics, Ludwig–Maximilians–Universität, D-80799 München,
   Germany, email helmut@stat.uni-muenchen.de
[2] Biostatistical Centre, Catholic University Leuven, Belgium

Abstract: We propose a general method for handling misclassification in the
context of regression models. The simex procedure form the theory of models
with continuous measurement error is applied to misclassification. The basic idea
is to fit a model for the relationship between the amount of misclassification and
the estimators of the parameters of interest by simulation. In the second step
this model is used for extrapolating back to the case of no misclassification. We
describe the procedure and given an example from a study on dental health in
Belgium.

Keywords: Misclassification; Simex; Logistic Regression; Response Error

1   Introduction

In general regression problems covariates and responses are often measured
with random error. In the case of discrete variables the measurement error
is referred as misclassification. While measurement error models have re-
ceived much attention in the literature there are only few recent papers on
misclassification.
We develop a new general approach for handling misclassification in discrete
covariates or responses in regression models. The simulation and extrapo-
lation (SIMEX) method (Cook and Stefansiki (1995)), which was originally
designed for handling additive covariate measurement error, is transfered
to the case of misclassification. The statistical model for characterizing mis-
classification is given by the transition matrix Π from true to the observed
variable. We exploit the relationship between the size of misclassification
and bias in estimating the parameters of interest. Assuming that Π is known
or can be estimated from validation data we simulate data with higher mis-
classification and extrapolate back to the case of no misclassification.

2   The procedure

We refer to a general regression problem with response Y and with a dis-
crete regressor X and further correctly specified regressors Z, where β is

the parameter of interest. We denote the possibly misspecified variable by X^* for the corresponding correctly measured (gold standard) variable X. Usually misclassification error is characterized by the misclassification matrix Π, which is defined by its components

$$\pi_{ij} = P(X^* = i | X = j).$$

Π is $k \times k$ matrix , where k is number of possible outcomes for X. If misclassification error is ignored, the corresponding estimator of β is called the naive estimator $\hat{\beta}_{na}$. The probability limit of the naive estimator is denoted by β^*. The existence of β^* and its determination can be done by the theory of misspecified models, see e. g. White (1982) . It depends on the model and on the misclassification matrix, i.e. $\beta^* = \beta^*(\Pi)$. We assume $\beta^*(Id_{k \times k}) = \beta$, i. e. that the estimator is consistent if no misclassification is present. We define the function

$$
\begin{aligned}
\lambda &\longrightarrow \beta^*(\Pi^\lambda) \\
\Pi^\lambda &:= E\Lambda^\lambda E^{-1}
\end{aligned}
\tag{1}
$$

where Λ is the diagonal matrix of eigenvalues and E is the matrix of the relating eigenvectors. The reason for analyzing (1) is that it is possible to simulate data with higher misclassification: If X^* has misclassification Π in relation to X and the Vector X^{**} is related to X^* by the misclassification matrix Π^λ then X^{**} is related to X by the misclassification matrix $\Pi^{\lambda+1}$. This is true if the two misclassification mechanisms are independent. One example is the logistic regression model with a binary misclassified covariate. It turns out, that function (1) can be well approximated by a log linear or a quadratic parametric function, i.e.

$$\lambda \longrightarrow \beta^*(\Pi^\lambda) \approx \mathcal{G}(\lambda, \Gamma) \tag{2}$$

The misclassification SIMEX procedure is as follows. Given data $(Y_i, X_i^*, Z_i)_{i=1}^n$ we denote the naive estimator by $\hat{\beta}_{na}[(Y_i, X_i^*, Z_i)_{i=1}^n]$.

1. Simulation step

For a fixed grid of positive values $\lambda_1 \ldots \lambda_m$ we simulate B new pseudo data sets by

$$X_{b,i}^*(\lambda_k) := MC[\Pi_k^\lambda](X_i^*), \quad i = 1, \ldots, n; \ b = 1, \ldots B; k = 1, \ldots, m. \tag{3}$$

where $MC[M](X*_i)$ denotes the simulation of a variable out of X_i^* with misclassification matrix M. Then we define $\lambda_0 = 0$, $\hat{\beta}(\lambda_0) = \hat{\beta}_{na}[(Y_i, X_i, Z_i)_{i=1}^n]$ and

$$\hat{\beta}(\lambda_k) := B^{-1} \sum_{b=1}^{B} \hat{\beta}_{na} \left[(Y_i, X_{b,i}^*(\lambda_k), Z_i)_{i=1}^n\right], k = 1, \ldots m. \tag{4}$$

The mean in (4) can be replaced by the median, if there are problems with stability.

2. Extrapolation step

Note that $\beta(\lambda_k)$ is an average over naive estimators corresponding to data with misclassification matrix $\Pi^{1+\lambda_k}$. So a parametric model $\mathcal{G}(\lambda, \Gamma)$ is fitted by least squares to $[\lambda_k + 1, \hat{\beta}(\lambda_k)]_{k=0}^m$, yielding an estimator $\hat{\Gamma}$. Then the MC-SIMEX estimator is then given by

$$\hat{\beta}_{SIMEX} := \mathcal{G}(0, \hat{\Gamma}). \tag{5}$$

If β is a parameter vector, the SIMEX estimator can be applied to every component of β separately like the original SIMEX. The application of the SIMEX for a misclassified variable Y is defined in the same way. In the simulation we have to simulate pseudo data $Y_{i,b}^*(\lambda_k)$.

The estimator $\hat{\beta}_{SIMEX}$ is consistent if the extrapolation function is correctly specified, i. e. $\beta^*(\Pi^\lambda) = \mathcal{G}(\lambda, \Gamma)$. for some parameter vector Γ. Usually this is not the case, but if $\mathcal{G}(\lambda, \Gamma)$. is a good approximation of $\beta^*(\Pi^\lambda)$ then approximate consistence will hold. To find suitable candidate for the function $\mathcal{G}(\lambda, \Gamma)$ we present the relationship between β^* and the misclassification parameter λ for some special cases. An example is given in Figure 1 for the case of logistic regression with one misclassified covariate.

The procedure can be generalized for misclassified responses and even for more than one misclassified regressor.

3   Application to the Caries study

The Signal-Tandmobiel study is a 6 year longitudinal oral health study involving 4468 children conducted in Flanders (Belgium). Data were collected on oral hygiene, gingival condition, dental trauma, prevalence and extent of enamel developmental defects, fluorosis, tooth decay, presence of restorations, missing teeth, stage of tooth eruption and orthodontic treatment need, all using established criteria. The children were examined annually for a period of six years (1996-2001). Our response of interest is the *dmf*, a binary variable equal to 1 if the tooth is decayed (d), missing due to caries (m) and filled (f) teeth, and 0 otherwise. The data were done by different examiners. In a calibration exercise it turned out that there was considerable misclassification in the data. The effect and correction for misclassification has been done for this study at one time point, see Mwalili et al.(2004).

We present a longitudinal analysis using GEE for four teeth, that is the first molars. Our main regressor variables are x- & y-coordinates of the schools of the children accounting for a possible spatial effect, age and gender. We also have tooth dummies and their possible interaction terms in our model. This model was fitted using GEE (PROC GENMOD of SAS) with MC-SIMEX correction for misclassification using log linear and quadratic extrapolation. The correction was done in two ways: (1) using a pooled misclassification matrix for all examiners and (2) using a misclassification matrix for each examiners.
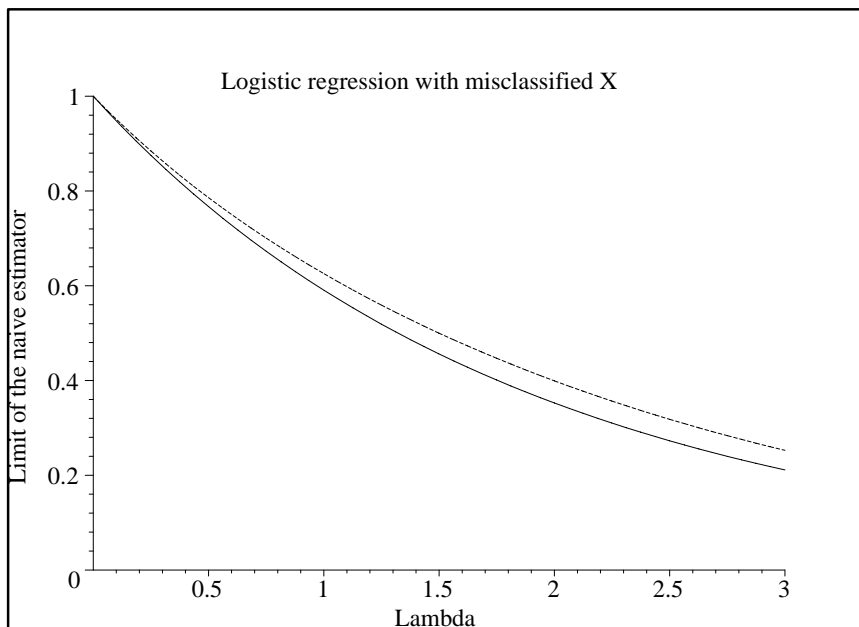
FIGURE 1. Limit of the naive estimator in the logistic Model $Y = \beta_0 + \beta_1 X$ with binary misclassified X. Here, $\beta_1 = 1$ and $\beta_0 = -2.\pi_{00} = \pi_{11} = 0.8$ (solid line) $\pi_{00} = 0.9, \pi_{11} = 0.7$ (dashed line), $\pi_{00} = 0.7, \pi_{11} = 0.9$ (dotted line).

The corrected parameter estimates were all larger than the naive estimates. Thereby adjusting for the attenuation effect due to misclassification of the dmf-score. The adjustment using different misclassification matrix for each examiner gives relatively less point estimates than the adjustment with a single fixed misclassification matrix for all examiners.

We discuss the variance estimation and taking into account that the misclassification matrix is only estimated with rather low precision. Furthermore we present a simulation study which gives good results for the MC-SIMEX procedure in particular for the log linear extrapolation function.

References

Cook, J. and Stefanski, L. (1995). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, **90**, 1314–1328.

Mwalili, S., Lesaffre, E.and Declerck, D. (2004). A bayesian ordinal logistic regression model to correct for inter-observer measurement error

in a geographical oral health study. *Journal of the Royal Statistics Society, Series C*, to appear.

White, H. (1982).Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–25.

Statistical inference for data files that are computer linked

Brunero Liseo [1], Andrea Tancredi [1]

[1] Dipartimento di Studi Geoeconomici, Statistici e Storici per l'Analisi Regionale, Università di Roma "La Sapienza"

Abstract: Record linkage refers to the use of an algorithmic technique to match records from different data sets that correspond to the same statistical unit, but lack unique personal identification code. In general, the margin of two (or more) data-bases can be important for two reasons. Firstly, *per sé*, i.e. to obtain a larger and richer data-file. Secondly to perform subsequent statistical analyses, based on information which is not simultaneously present in both files. In this paper we will propose a Bayesian approach particularly suitable in the latter case

1 Introduction

The need of record linkage (RL) techniques is steadily increasing in various chapters of statistics. For example, in official statistics record linkage is a preliminary step when the size of a population is estimated via capture-recapture techniques, especially when the target population is elusive (non regular immigrants in European Community are an example) and differences in identification variables in the two occasions are frequent. The creation of integrated data bases obtained by the merging of existing one is also important in epidemiology where RL is commonly used in cohort studies to ascertain the study outcome and, as such, its accuracy in classifying the outcome can be described using the standard epidemiological terms of sensitivity and positive predictive value. In general, the margin of two (or more) data-bases can be important both *per sé*, i.e. to obtain a larger and richer data-file and to perform subsequent statistical analyses, based on information which is not simultaneously present in both files. To give an example of the latter, suppose we have two computer files $\mathcal{A}$ and $\mathcal{B}$ whose records relate respectively to units of partially overlapping populations $\mathcal{P}_\mathcal{A}$ and $\mathcal{P}_\mathcal{B}$. The two files consist of several fields, or variables, either quantitative or qualitative. The objective of record linkage is to find all the pairs of units (a,b), $a \in \mathcal{A}$ and $b \in \mathcal{B}$, such that a and b refer actually to the same unit. Suppose that the observed variables in $\mathcal{A}$ are $(Z, W_1, W_2, \cdots, W_k)$ while in $\mathcal{B}$ we observe $(W_1, W_2, \cdots, W_k, X)$. Then we might be interested in studying a linear regression analysis between Z

and X, restricted to those couple of records which we declare as matches. The intrinsic difficulties present in such a simple problem are discussed in Scheuren and Winkler (1993) and Lahiri and Larsen (2004).

In a more general framework, suppose that file $\mathcal{A}$ contains the variables $(Z, W_A) = (Z_1, Z_2, \cdots Z_h, W_1, W_2, \cdots, W_k)$ observed on ν_A units, while $\mathcal{B}$ contains the variables $(W_B, X) = (W_1, W_2, \cdots, W_k, X_1, X_2, X_p)$; our goal is to use the key variables $(W_1, W_2, \cdots, W_k)$ to detect the true links between X_A and X_B and to perform a statistical analysis involving vectors Z and X restricted to those records which have been defined matches. To perform this task, we present a fully Bayesian approach which is particularly suitable to accomplish the above desideratum. Under our approach all the uncertainty about the matching process is retained in the subsequent inferential steps. Our approach can be considered an improvement and a generalization of the Bayesian model described in Fortini *et al.*. We will present the general theory underlying the model and illustrate its performance with a linear regression model.

2   Bayesian Record Linkage

2.1   The usual statistical model for record linkage

We first examine the classical approach to the record linkage problem, see Jaro (1989), Larsen and Rubin (2001). Consider two data files $\mathcal{A}$ and $\mathcal{B}$, with respectively ν_A and ν_B units. Let us call A and B the two sets (lists) of observed units, $a = 1, ..., \nu_A$, $b = 1, ..., \nu_B$. We assume that at least some units are present in both lists. The set of all ordered pairs $A \times B = \{(a, b) : a \in A, b \in B\}$ can be splitted into $\mathcal{M} = \{(a, b) \in A \times B : a = b\}$ the set of matches, and $\mathcal{U} = \{(a, b) \in A \times B : a \neq b\}$ the set of non-matches. In order to decide whether a pair (a, b) is in $\mathcal{M}$ or $\mathcal{U}$, we may compare variables observed in both the files (e.g. surname, name, sex, address, etc. for individuals). Let us assume we have k key variables, $k \geq 1$, whose observations in the two data lists are denoted by: $w_a = (w_{a,1}, w_{a,2}, ..., w_{a,k})$, $a \in A$, and $w_b = (w_{b,1}, w_{b,2}, ..., w_{b,k})$, $b \in B$. In general, the comparison y_{ab} of the key variables between two units $a \in A$ and $b \in B$ will be a function of w_a and w_b. One commonly assumed comparison function is a vector of k elements, $y_{ab} = (y_{ab}^1, ..., y_{ab}^k)$ with $y_{ab}^h = 1$ if $w_{a,h} = w_{b,h}$ and 0 otherwise for $h = 1, \ldots, k$ In this case the comparison vector y_{ab} can assume 2^k different values which we will indicate with y_i where $i = 1, \ldots, 2^k$. In order to decide whether a pair (a, b) with comparison vector y_{ab} should be linked or not, Fellegi and Sunter suggest to consider the sampling distribution of the comparison vectors in $\mathcal{M}$, say $m(y)$, and the corresponding distribution in $\mathcal{U}$, $u(y)$. The decision rule for the pair (a, b) is based on the likelihood ratio $t(y_{ab}) = \frac{m(y_{ab})}{u(y_{ab})}$. Fellegi and Sunter (1969) discuss several frequentist optimality properties

of such decision rule. Given that neither $m(y)$ nor $u(y)$ are known, most of the literature on record linkage concentrates on how to estimate them. The usual assumptions are that both the status of a pair (let's say c_{ab}, where $c_{ab} = 1$ when a pair (a, b) is a true match and 0 otherwise) and the comparison vector Y are random variables. Also, a general latent structure is assumed via the configuration matrix $c = \{c_{ab}, a \in A, b \in B\}$, so that the values c_{ab}, $(a, b) \in A \times B$, are assumed to be i.i.d. Bernoulli r.v. such that for all a, b, $P(c_{ab} = 1) = p$; the comparison vectors Y_{ab}, $(a, b) \in A \times B$, are assumed to be i.i.d. replications of the r.v. Y whose distribution has the the the mixture structure $P(Y = y) = p\, m(y) + (1 - p)\, u(y)$; finally the random vectors (c_{ab}, Y_{ab}), $(a, b) \in A \times B$, are i.i.d. with distribution given by $P(c = c,\, Y = y) = \big(p\, m(y)\big)^c \big((1 - p)\, u(y)\big)^{1-c}$, with $c = 0, 1$.

2.2   The Bayesian model

The Bayesian model comprises the prior distribution on the unknown parameters and the conditional distribution of the observed data given the unknown parameters. The observed data are given by the vector $y = (y_{11}, \ldots, y_{\nu_a, \nu_b})$ while the unknown parameters are the matrix c, the vector $m = (m_1 \ldots, m_{2^k})$ where $m_i = P(Y_{ab} = y_i | c_{ab} = 1)$ and the vector $u = (u_1 \ldots, u_{2^k})$ where $u_i = P(Y_{ab} = y_i | c_{ab} = 0)$. The conditional distribution of the observed vector y given c, m, u is

$$
f(y|c, m, u) \quad = \quad \prod_{a=1}^{\nu_A} \prod_{b=1}^{\nu_B} \left[\prod_{i=1}^{2^k} m_i^{d(y_{ab}, y_i)} \right]^{c_{ab}} \left[\prod_{i=1}^{2^k} u_i^{d(y_{ab}, y_i)} \right]^{1 - c_{ab}}
$$

where $d(y_{ab}, y_i) = 1$ if $y_{ab} = y_i$ and 0 otherwise. In what follows, we will assume that m and u are a priori independent on c. We take a Dirichlet distribution as a prior distribution both for m and for u. In particular $u \sim \mathcal{D}(\alpha_1, \ldots, \alpha_{2^k})$ and $u \sim \mathcal{D}(\beta_1, \ldots, \beta_{2^k})$ where $\log \alpha_i = (\sum_{i=1}^{k} y_i^k - \phi) \log \theta$ and $\log \beta_i = (\phi - \sum_{i=1}^{k} y_i^k) \log \theta$. Fortini *et al.* show how to calibrate the hyperparameters θ and ϕ. To complete the model we need to give a prior distribution to the matrix c. Let c be a matrix such that where $c_{ab} \in \{0, 1\}$, $\sum_{a=1}^{\nu_A} c_{ab} \le 1$ $\sum_{b=1}^{\nu_B} c_{ab} \le 1$. Let $t = \sum_{ab} c_{ab}$ be number of matches, let $T_m = \min\{\nu_A, \nu_B\}$ be the maximum number of matches and let $T_q = \max\{\nu_A, \nu_B\}$. The prior distribution on c is built in two stages. In the first stage we assume that t, the number of matches, has binomial distribution with paramters ξ and T_m In the second stage we assume a uniform distribution on the space of all possible matrices with t matches. Notice that the hyperparameter ξ represents the probability that a generic unit in the smaller file belongs to the bigger file. We can consider ξ either known or unknown. In the latter case a Beta prior can be used. Moreover we observe that $E(c_{ab}) = p$ where $p = \xi/T_q$. Then p represents the probability that a generic couple (a, b) is a match. The Bayesian model proposed

in this paper is too complex to be amenable to analytical calculations. Hence, we shall use MCMC methods, and in particular a Gibbs sample algorithm. In fact we are able to produce random variates from each of the full conditionals of the model.

3   A general approach for dependent data

In this section we discuss the problem of the statistical modelling for multivariate observations obtained by RL techniques. Considering the posterior distribution for the matrix c produced by the Bayesian procedure described above we obtain a point estimate for c that can be used for the subsequent inference. However, in this case we do not take account of record linkage uncertainty and we risk to overestimate the precision of the estimates. To overcome this problem we propose the following model. Let $D = (y, z, x) = (y_{11} \ldots, y_{\nu_A \nu_B}, z_1, \ldots, z_{\nu_A}, x_1, \ldots, x_{\nu_B})$ be the available data where y_{ab} is the comparison vector for the units a and b, z_a is the value of the variable Z observed on unit a of the file A and x_b is the value of the variable X observed on unit b of the file B. We indicate with

$$p(y, z, x | c, m, u, \theta) = p(y | c, m, u, \theta) p(x, z | c, y, m, u, \theta) \tag{1}$$

the general statistical model for suck kind of data. The quantities c, m, u are the record linkage parameters while θ represents the parameter vector of the joint distribution (X, Z). It is reasonable to assume that given the matrix c, the comparisons y do not depend on θ. Moreover we can assume that, given the matrix c, the law of (X, Z) depends neither on the observed comparison vectors y nor on the parameters of the comparison vectors m and u. In this way we write the model (1) as

$$p(y | c, m, u) p(x, z | c, \theta). \tag{2}$$

where the first term is the usual likelihood for the RL model while the second term depends on the dependence structure between x and z. Conducting inference for θ by the model (2) we take account of the RL uncertainty and at the same time we improve the RL procedure by the information provided by the statistical relationship between the variables z and x.

3.1   Regression analysis

We now face the problem of the regression analysis with linked data. Suppose we have two variables Z, X where the marginal density of Z is $f_Z(z)$ and Z given $X = x$ is normal distributed with density $\phi(z; x\beta, \sigma_{z|x})$. For the moment we assume that β, $f_Z(z)$ and $\sigma_{z|x}$ are known. On file $\mathcal{A}$ we observe the variable Z while in file $\mathcal{B}$ we observe the variable X. The likelihood ratio

$$R = \frac{P(z_a, x_b | (a, b) \in \mathcal{M})}{P(z_a, x_b | (a, b) \in \mathcal{U})} = \frac{P(z_a | x_b, (a, b) \in \mathcal{M})}{P(z_a | x_b, (a, b) \in \mathcal{U})} = \frac{\phi(z_a; x_b \beta, \sigma_{z|x})}{f_Z(z_a)}$$

will provide useful information for the matching process. In fact given a unit $a \in \mathcal{A}$ we expect higher values of R when the record b produces a value of $x_b \beta$ simalar to z_a (which is the case when the pair (a, b) is actually a match) and small values for R otherwise. Let z be the vector $(z_1, \ldots, z_{\nu_A})$ and let x be the vector $(x_1, \ldots, x_{\nu_A})$. In such a situation we will assume

$$p(z|c, x) = \prod_{a=1}^{\nu_A} \prod_{b=1}^{\nu_B} \phi(z_a; \beta x_b, \sigma_{z|x})^{c_{ab}} \prod_{a=1}^{\nu_A} f_Z(z_a)^{1 - \sum_{l=1}^{\nu_B} c_{al}}.$$

Moreover assuming, as in the general framework, that the comparison vectors y and z are independent given the matrix c and that y is independent on x given c we have $p(y, z|c, x, m, u) = p(y|c, m, u)p(z|c, x)$. We may show by simulation that, the use of the information given by the linear relationship between Z and X with the model $p(y, z|c, x, m, u)$, improves the matching process. Finally we observe that when β is unknown we can easily produce posterior estimates. It is enough to modify the Gibbs algorithm adding a simulation step from the conditional posterior distribution for β. In fact given the matrix c the conditional posterior distribution for β is obtained considering the pairs (a, b) such that $c_{ab} = 1$ as true matches. In this way, estimating β with the marginal posterior mean, we will automatically take account of the matching process uncertainty and this can be worthwile when the regression step is the primary goal of the analysis. In this context we compare our approach with the frequentist proposals of Lahiri and Larsen (2004).

References

Fellegi, I., and Sunter, A. (1969). A theory of record linkage. *Journal of the American Statistical Association*, **64**, 1183-1210.

Fortini, M., Liseo, B. Nuccitelli, A. and Scanu, M. (2001). On Bayesian record linkage. *Research in Official Statistics.* 185-198.

Jaro, M. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, **84**, 414-420

Lahiri, P. and Larsen, M. (2004). Regression analysis with linked data. *Journal of the American Statistical Association, to appear.*

Larsen, M., and Rubin, D. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, **96**, 32-41

Scheuren, F., and Winkler,W. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, **19**, 39-58.

Advances in Covariance Modelling

Gilbert MacKenzie[1]

[1] Centre for Medical Statistics, Keele University, Keele, Bedfordshire ST5 5BG, UK. e-mail: g.mackenzie@keele.ac.uk

Abstract

Conventionally, in longitudinal studies, the mean structure has been thought to be more important than the covariance structure between the repeated measures on the same individual. Often, it has been argued that, with respect to the mean, the covariance was merely a 'nuisance parameter' and, consequently, was not of 'scientific interest'. Today, however, one can see that from a formal statistical standpoint, the inferential problem is entirely symmetric in both parameters. In recent years there has been a steady stream of new results and we pause to review some key advances in the expanding field of covariance modelling, In particular, developments since the seminal work by Pourahmadi (1999, 2000) are traced. While the main focus is on longitudinal data with continuous responses, emerging approaches to joint mean-covariance modelling in the GEE, and GLMM arenas are also considered briefly.

Keywords Cholesky Decomposition, Covariance Modelling, Joint Model Space, Longitudinal Studies, GEE, GLMMs.

1 Introduction

The conventional approach to modelling longitudinal data places considerable emphasis on estimation of the mean structure and less on the covariance structure, between repeated measurements on the same subject. Often, the covariance structure is thought to be a of secondary scientific interest and is selected from a limited menu of structures, e.g., compound-symmetry, AR(1), AR(2) or a saturated model.

However, from a formal statistical standpoint the inferential problem is entirely symmetric in both parameters μ and Σ. We note that it was (Rao, 1965), who first showed that the mean is covariance invariant, only when the covariance matrix belongs to a special class of covariance structures - Rao's Simple Structure. When Σ is outwith this class one may anticipate that a suboptimal choice of Σ may influence μ and *vice versa*. If so, one approach

is to search the joint model space, $\{\mathcal{M} \times \mathcal{C}\}$, in order to determine the optimal estimators $(\hat{\mu}, \hat{\Sigma})$. The concept of the joint model space is central to what follows.

Determining the structure of Σ, from the data, rather than from a pre-specified menu, may at first seem daunting, whence the idea of searching the entire model space, $\{\mathcal{C}\}$, for Σ, may seem prohibitive. The final demand, that one conduct a simultaneous search of the Cartesian product $\{\mathcal{M} \times \mathcal{C}\}$ may seem impossible. However, these apparently difficult tasks can be accomplished easily for a particular, but very general, class of covariance structures, $\{\mathcal{C}^*\}$, defined below.

2   Covariance Modelling

2.1   Rationale

It is well known that in the linear model, applied to longitudinal studies, the maximum likelihood estimates of the regression coefficients, take the Weighted Least Squares (WLS) form:

$$\hat{\beta}_\Sigma = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y \tag{1}$$

where the dependence of β on Σ has been emphasized. In practice this dependence is often ignored. The usual approach is to adopt a two-stage model selection strategy, fixing the structure of Σ first and then finding the maximum likelihood estimates of $\hat{\beta}$ and $\hat{\Sigma}$ in simultaneous *estimation*. This, may be joint *estimation*, but it is not joint (mean-covariance) *model selection*, because a search of the joint mean-covariance space, $\{\mathcal{M} \times \mathcal{C}\}$, has not been conducted.

Perhaps such a search is not necessary. One might conjecture that $\hat{\beta}$ is Σ invariant. However, this is hardly compelling in view of the form of (1), in which Σ^{-1} clearly acts as a weight matrix. Thus, if Σ is not the truth, one should expect the magnitude of the fixed effects to be distorted by an amount which is a function of the dis-similarity between Σ and the true variance-covariance matrix.

A natural first question is to enquire whether there is any situation in which $\hat{\beta}$ is Σ invariant? One obvious case arises when $\Sigma \equiv I$, i.e., when the errors are i.i.d.. More importantly, Rao (1965) showed that $\hat{\beta}$ is Σ invariant when

$$\Sigma = X\Gamma X' + Q\Theta Q' \tag{2}$$

where Γ of order $(p \times p)$ and Θ of order $((p-m) \times (p-m))$ are positive definite and Q is a $(p \times (p-m))$ matrix orthogonal to X, i.e., $Q'X = 0$. In this formulation there are exactly m repeated measurements over time.

This result shows that $\hat{\beta}$ is *not* Σ invariant, in general, but only when Σ lies in Rao's Simple Covariance Structure (SCS) defined by (2). The next natural question is which of the commonly occurring covariance structures

utilized in longitudinal modelling lies in SCS? The answer to this question is largely open, although it may be shown that compound symmetry (CS) is contained in SCS, but that for example AR(1) is not (Pan & Fan, 2002).

The foregoing has highlighted the impact of covariance mis-specification on $\hat{\beta}$, mainly, because this issue is not widely understood. However, such mis-specification may also impact on the standard error of $\hat{\beta}$. Thus, the next question is how then can current practice be improved?

2.2   Joint Regression Model

In the context of a longitudinal study with a Gaussian response, the solution is based on a modified Cholesky decomposition of the usual marginal covariance matrix $\Sigma(t, \theta)$, where t represents time and θ is a low-dimensional vector of parameters describing dependence on time. The decomposition leads to a reparametrization, $\Sigma(t, \varsigma, \phi)$, in which the new parameters have an obvious statistical interpretation in terms of the natural logarithms of the innovation variances, ς, and generalized autoregressive coefficients, ϕ, Pourahmadi (1999, 2000). These unconstrained parameters are modelled, parsimoniously, as different polynomial functions of time

$$\mu_{ij} = x'_{ij}\beta \qquad \phi_{ijk} = z'_{ijk}\gamma \qquad \varsigma_{ij} = h'_{ij}\lambda \qquad (3)$$

where a polynomial representation for the mean structure has been included in order to fit a joint mean covariance model. Here, β, γ and λ are the three regression parameters of primary scientific interest while z and h are particular polynomials in lag and time, respectively.

2.3   Covariance Classes

The covariance class $\{\mathcal{C}^*\}$ defined by the last two polynomial regressions in (3) is capable of representing a wide variety of stationary and non-stationary covariance structures and provides a relatively smooth method of transition from structure to structure, compared with relatively limited menu selection methods. An additional point to consider is that in $\{\mathcal{C}^*\}$ the transformed covariance parameters now have an interpretation which is relatively unfamiliar to bio-statisticians, but which is used routinely in time series and Kalman filtering applications (MacKenzie & Reeves, 2002). Of course, $\{\mathcal{C}^*\}$, is not the only type of regression-based covariance class which may be defined at (3). Smoother, non-parametric, regression models may be preferred to enrich the class and these are being developed.

2.4   Optimal Mean-Covariance Modelling

The optimal joint-mean covariance model may be found by a direct search of $\{\mathcal{M} \text{ x } \mathcal{C}^*\}$. This amounts to determining the degrees, (p, q, d), of the

three polynomial functions in (3) which minimize some suitable model selection criterion such as AIC or BIC, over the joint model space. When the longitudinal data are balanced with m repeated measurements, $\{\mathcal{M} \text{ x } \mathcal{C}^*\}$ is a m-cube. Pan & MacKenzie (2003) show how to search $\{\mathcal{M} \text{ x } \mathcal{C}^*\}$ efficiently using a profile BIC-based algorithm. The optimum degree triple (p_c^*, q_c^*, d_c^*) is found as

$$
\begin{aligned}
p_c^* &= \arg\min_p\{\text{BIC}(p,s,s)\} \quad q_c^* = \arg\min_q\{\text{BIC}(s,q,s)\} \qquad (4) \\
d_c^* &= \arg\min_d\{\text{BIC}(s,s,d)\}
\end{aligned}
$$

where s stands for saturated degree. The profile BIC algorithm linearizes the search.

2.5   Modelling Heterogeneity

An important application of these regression methods occurs in longitudinal randomized controlled trials. Conventionally, it is assumed that the intervention will influence the evolution of the mean, but it is presumed that it will *not* influence the covariance structure. This asymmetrical approach to modelling the mean and covariance pervades much statistical practice. With hindsight, this is simply one model choice and in many cases it may be untenable. Equations (3), however, now render it a *testable* model choice, by enabling one to include the treatment indicator and treatment by time interactions in the last two equations of the model. MacKenzie & Pan (2001) illustrated the method of analysis using Kenward's (1987) cattle data, demonstrating *inter alia* that intervention had altered the covariance structure, an effect which was missed in the original analysis. The above procedure models the covariance structure in terms of *fixed effects* which may be different in the mean and covariance structures.

2.6   Modelling Conditional Covariance

For the linear mixed model, Laird & Ware (1982) showed that the marginal covariance matrix may be decomposed as

$$\Sigma = \Sigma_B(t;\theta_B) + \Sigma_W(t;\theta_W) \qquad (5)$$

where $\Sigma_B(t;\theta_B)$ represents the between subject covariance while $\Sigma_W(t;\theta_W)$ represents the within subject covariance and θ_B and θ_W are low dimensional vectors describing their respective dependencies on time. In some parametrizations $\Sigma_B(t;\theta_B)$ may not depend on time, but may depend on stationary covariates, as in the previous section. Classically, here, there are two covariance menus to be recursed. However, the regression modelling approach can, most obviously, be applied to $\Sigma_W(t;\theta_W)$, given an agreed structure for $\Sigma_B(t;\theta_B)$. Pan & MacKenzie (2001) used the E-M algorithm to obtain a *data driven* estimate of $\Sigma_W(t;\theta_W)$.

2.7   GEEs & GLMMs

The modelling strategy outlined above assumes a Gaussian response. However, Ye and Pan (2003) exploit the GEE framework to propose three estimating equations for joint mean-covariance models involving continuous responses (not necessarily Gaussian). They also studied hypothesis tests for parameters involved in the mean, the autoregressive coefficients and the innovation variances, using score-type tests. Moreover, they have investigated the asymptotic properties of the parameter estimates obtained.

In further work, Pan *et al* (2004) have extended their procedures to modelling covariance structures in the GLMM framework. The approach differs from that outlined above as the modelling is conducted in the latent, rather than in the observation, space.

3   Discussion

Covariance regression modelling is now a substantive area of statistical modelling. As a field, it has been developing steadily and an increasing range of versatile techniques, including Bayesian methods (Daniels and Pourahmadi, 2002), have become available in the last five years. It is too soon, of course, to claim that of all the outstanding problems have been solved. This is simply not true, but considerable progress has been made and more is expected in the years ahead.

References

Kenward, M. G. (1987). A method for comparing profiles of repeated measurements. *Appl. Statist.* **36**, 296-308.

MacKenzie, G., Pan, J. (2001). Modelling marginal covariance structures in linear mixed models. In: *Proceedings of the 16th IWMS*, 275-282, Odense (DK).

Pan JX & MacKenzie G. (2003). On modelling mean-covariance structures in longitudinal studies *Biometrika*, **90**, 239-244.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, **86**, 677–690.

Pourahmadi, M. (2000). MLE of generalised linear models for multivariate normal covariance matrix. *Biometrika*, **87**, 425–435.

Rao, C.R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, **52**, 447–458.

Nonparametric Modelling of Longitudinal Data: A Varying Coefficients Model

Susan Orbe-Mandaluniz[1], Vicente Núñez-Antón[2], and Juan M. Rodríguez-Póo[3]

[1] Departamento de Econometría y Estadística, Universidad del País Vasco, Avenida del Lehendakari Aguirre 83, E-48015 Bilbao, Spain, e-mail: `etpormas@bs.ehu.es`
[2] Departamento de Econometría y Estadística, Universidad del País Vasco, Avenida del Lehendakari Aguirre 83, E-48015 Bilbao, Spain
[3] Departamento de Análisis Económico, Universidad de Zaragoza, Gran Via 2, E-50005 Zaragoza, Spain

Abstract: We propose a very flexible and general semiparametric model that allows for time-varying coefficients and/or covariate-varying (including groups-varying or subjects-varying) coefficients in a longitudinal data setting. Tests for model specification are proposed and, thus, this proposal allows to discriminate between the different sources of variation for the regression coefficients. The model is applied to several longitudinal data examples and, in addition, its performance is studied when compared to other more restricted proposals.

Keywords: Nonparametric kernel estimation; Varying coefficients; Unstructured covariances; Model specification tests; Longitudinal data.

1   Introduction and General Model.

The analysis of longitudinal data, where experimental units (normally allocated to different treatments or groups) are measured over a period of time, has been studied extensively. In particular, it is interesting to separate what is common to the whole population from what is specific to each treatment or group, and also from what is specific to each individual. These notions have been previously analyzed in a parametric setting by Diggle et al. (1994), among others. Núñez-Antón and Zimmerman (2000) and Zimmerman and Núñez-Antón (2001) have analyzed these notions for several data sets and proposed a joint mean and covariance analysis for modelling these structures. Thus, it is of interest for researchers to be able to separate the different effects and the way they can depend on the covariates. For example, for the cattle data (see Kenward, 1987), a designed experiment in which cows receiving two treatments for intestinal parasites were weighted over time, or for the dogs data (See, Grizzle and Allen, 1969) a designed experiment in which measurements of coronary sinus potassium concentration after occlusion on four groups of dogs were taken over time,

researchers were interested in separating the common, group (i.e. type of treatment) and individual effects believed to be present in these data sets. In addition, it may also be of interest to study the possible dependence of any of these effects on time and the different features in the within subjects covariance structure. These two ideas and the steps to carry them out can, of course, provide a clear picture of the main properties of the dependency between the response variable and time and/or covariates for these data sets. Profile plots for both data sets, and additional data sets we have considered, indicate that it is quite hard to reckon a precise parametric form to use for the model in these data sets.

Many parametric models could be used to estimate the separate effects in the balanced data case (see for example, Potthoff and Roy, 1964; among others). However, when the data are unbalanced (if the model is linear see, for example, Longford, 1993), and when the models are not necessarily assumed to be linear, it is possible to fit each curve individually and to work on the parameter set afterward (Caussinus and Ferré, 1992) to investigate the relations and differences between the subjects. In order to estimate common and specific effects Laird and Ware (1982), in the linear case, and Lindstrom and Bates (1990), in the nonlinear one, used mixed effects models in which the common part is the fixed effect while the specific ones are the random effects of the model. Then, maximum likelihood estimators are obtained from the EM algorithm. This requires a clear knowledge of what is common and what is specific because in practical situations it is very important to decide which parametric model to assume.

Another added difficulty is the parametric specification for the within-subjects covariance structure. Most of the different proposed approaches allowed for unbalanced data and were applied to completely specified linear models. In addition, they were also able to investigate the effect of the groups, but could not separate (i.e. distinguish) the three components present in these data sets and the possible dependence they may have on time. Therefore, a model for this data set must be able to include: (i) a common component, representing the fact that individuals come from the same population, (ii) a group component, since there are different treatments, (iii) an individual component, and (iv) a possible time-dependence for each of these different components.

Nonparametric approaches have been developed in order to avoid the difficulty of specifying a parametric model, by estimating the relationship between the response variable and time over a large class of smooth functions (see. e.g., Gasser et al., 1984). The main drawback of these models is that nonparametric regression estimates may behave quite poorly for small sample sizes and, unfortunately, this is quite often the case in practice. In order to partially solve this problem, and for the case where independence among measurements is assumed, a two-stage approach was developed by Boularan et al. (1994) to study the dependence between height and age. They used an additive model and were interested in estimating the com-

mon component and the group component (boys and girls). This two-stage approach would have the advantage that the mean part can be estimated very precisely by using the data on all m individuals, and it does allow individuals to be measured at different times. Even though this model allows for unbalanced data, it does not include all the components present in our data sets, and it does not take into account the within-subject covariance structure.

Therefore, there is a need to propose a model able to deal with unbalanced data and that allows us to estimate the three components present in the data and its possible dependence on time; it should also allow us to have general within subject covariance structures, and should be able to deal with the few observations usually available per subject. Along these lines, we consider a general linear varying coefficients model of the form:

$$Y_{ij} = \mathbf{X}_{ij}^{\mathrm{T}} \, \boldsymbol{\beta}_{ij} + \epsilon_{ij}, \tag{1}$$

where Y_{ij} represents the response at time t_{ij} for subject i $(i = 1, \ldots, m)$ at the j-th time $(j = 1, \ldots, n_i)$, $\mathbf{X}_{ij}$ denotes the $p \times 1$ vector of covariates for the i-th individual, that could include group dependence or time, $\boldsymbol{\beta}_{ij}$ is the $p \times 1$ vector of fixed and unknown parameters, that may depend on time and/or specific covariates, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{i,n_i})^{\mathrm{T}}$ is assumed to have zero mean and a full rank covariance matrix Ω_i. The coefficients $\boldsymbol{\beta}_{ij} = f_{ij}(t_{ij}, \mathbf{Z}_{ij})$ are determined by an unknown function $f_{ij}(\cdot)$ that can depend on time (i.e., through t_{ij}), and on individuals and/or on groups (i.e., through $\mathbf{Z}_{ij}$, where $\mathbf{Z}_{ij}$ usually includes a subset of the covariates included in $\mathbf{X}_{ij}$ that are not time-dependent). The model proposed in (1) represents a very flexible and general specification for most models in the sense that it allows for the estimation of the different effects and, in addition, allows the coefficients to vary with time and/or the covariates included in the model. Moreover, the flexibility of the proposed nonparametric estimation method allows the estimation of the coefficients without the need to specify the function $f_{ij}(\cdot)$, and the only requirement it has is the assumption of some degree of smoothness. The coefficients in (1) are not required to vary with the same covariates and, thus, we could consider models where a specific set of coefficients varies only with time $(\boldsymbol{\beta}_{ij} = f_{ij}(t_{ij}))$, whereas another set varies with given covariates included in $\mathbf{Z}_{ij}$. In fact, this model could very well study situations in which one wishes to assess at the same time the effect of a treatment over time and the effect of the treatment itself. Thus, the proposed model can be written as:

$$Y_{ij} = \left(\mathbf{X}_{ij}^{(1)}\right)^{\mathrm{T}} \boldsymbol{\beta}_{ij}^{(1)}(t_{ij}) + \left(\mathbf{X}_{ij}^{(2)}\right)^{\mathrm{T}} \, \boldsymbol{\beta}_{ij}^{(2)}(\mathbf{Z}_{ij}) + \epsilon_{ij}, \tag{2}$$

where, under some restrictions, $\mathbf{X}_{ij}^{(1)}$ represents the $p_1 \times 1$ vector containing the subset of non-time dependent covariates that go with the time dependent $p_1 \times 1$ vector of coefficients $\boldsymbol{\beta}_{ij}^{(1)}(t_{ij})$, and $\mathbf{X}_{ij}^{(2)}$ represents the

$p_2 \times 1$, $(p_1 + p_2 = p)$ vector containing the subset of covariates that go with the group or individual dependent $p_2 \times 1$ vector of coefficients $\boldsymbol{\beta}_{ij}^{(2)}(\mathbf{Z}_{ij})$. Special cases of model (2) include:

- If $\boldsymbol{\beta}_{ij}^{(2)}(\mathbf{Z}_{ij}) = \mathbf{0}$, no individual, group or other covariates effects are considered and, thus, the resulting model corresponds to the time-varying coefficient model proposed by Hoover et al. (1998).

- If there is no time effect; that is, if $\boldsymbol{\beta}_{ij}^{(2)}(t_{ij}) = \mathbf{0}$, we obtain a more general model than the one in Núñez-Antón et al. (1999) or Zeger and Diggle (1994).

In particular, the specification of model (1) allows for the possibility to test for the existence of each one of the components in (2) and, thus, the possibility of considering the general model (i.e., model (1)) or any of its special cases (i.e., model (2) or any of its two particular cases).

2   Data set and results

The proposed models were applied to the two data sets mentioned in Section 1, and the general conclusions indicate that:

- For the cattle data (see Kenward, 1987), there is a strong group difference and, thus, a group effect that changes over time. In addition, it is of interest to include an individual effect that may or may not change over time. Thus, the proposed model has to be the more general one (i.e., model (1)). These conclusions somehow agree with the ones previously obtained in the more restrictive models used by Zimmerman and Núñez-Antón (2001) and Kenward (1987).

- For the dogs data (see Grizzle and Allen, 1969), there is strong group difference that does not substantially changes over time. It is clear that group 1 (i.e., the control group) is significantly different from the rest. In addition, it is of interest to include an individual effect that may or may not change over time. Thus, the proposed model has to be the one that allows for separation between effects (i.e., model (2)).

In summary, the models proposed in Section 1, when applied to several examples in the context of longitudinal data, have shown to be very useful additions to the existing models and, given that they generalize earlier models, they represent a valuable way of testing for submodels, such as the ones described above or in the literature.

References

Boularan, J., Ferré, L. and Vieu, P. (1994). Growth curves: a two-stage nonparametric approach. *Journal of Statistical Planning and Inference*, **38**, 327-350.

Caussinus, H. and Ferré, L. (1992). Comparing the parameters of a model for several units by mean of principal components analysis. *Computational Statistics and Data Analysis*, **13**, 269-280.

Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

Gasser, T., Müller, H.G., Kölher, W., Molinari, L. and Prader, A. (1984). Nonparametric regression analysis of growth curves. *Annals of Statistics*, **12**, 210-219.

Grizzle, J.E. and Allen, D.M. (1969). Analysis of growth and dose response curves. *Biometrics*, **25**, 357-381.

Hoover, D.R., Rice, J.A., Wu, C.O. and Yang, L.P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**, 809-822.

Kenward, M.C. (1987). A method for comparing profiles of repeated measurements. *Applied Statistics*, **36**, 296-308.

Laird, N.M. and Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics*, **38**, 963-974.

Lindstrom, M.J. and Bates, D.M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, **46**, 673-687.

Núñez-Antón, V. and Zimmerman, D.L. (2000). Modelling nonstationary longitudinal data. *Biometrics*, **56(3)**, 699-705.

Núñez-Antón, V., Rodríguez-Póo, J.M. and Vieu, P. (1999). Longitudinal data with nonstationary error: a nonparametric three-stage approach. *Test*, **8**, 210-231.

Zeger S.L. and Diggle P.J. (1994). Semiparametric models for longitudinal data with applications to CD4 cell numbers in HIV seroconverters. *Biometrics*, **50**, 689-699.

Zimmerman, D.L. and Núñez-Antón, V. (2001). Parametric modelling of growth curve data: An overview (with discussion). *Test*, **10(1)**, 1-73.

Quasi-Monte Carlo Estimation in Generalized Linear Mixed Models

Jianxin Pan[1] and Robin Thompson[2]

[1] Department of Mathematics, The University of Manchester, Manchester M13 9PL, U.K. Email: jpan@maths.man.ac.uk
[2] IACR-Rothamsted and Roslin Institute, Harpenden AL5 2JJ, U.K. Email: robin.thompson@bbsrc.ac.uk

Abstract: Statistical inference for generalized linear mixed models (GLMM) is highly challenging because the marginalized likelihood may involve analytically intractable integrals. In this paper a Quasi-Monte Carlo (QMC) approach that generates integration nodes uniformly on a domain is proposed to approximate the maximum likelihood estimates (MLE). Theoretical issues are studied and numerical comparisons to existing procedures are made in terms of real-data analysis and simulation studies.

1 Generalized linear mixed models

Suppose y_i $(i = 1, 2..., n)$ are the responses. Let x_i and z_i be $(p \times 1)$ and $(q \times 1)$ covariate vectors associated with fixed effects β $(p \times 1)$ and random effects b $(q \times 1)$, respectively. Given the random effects b, the responses y_i are independent with means and variances:

$$E(y_i|b) = \mu_i \qquad \text{and} \qquad \text{var}(y_i|b) = \phi a_i^{-1} \nu(\mu_i) \qquad (1)$$

respectively, where ϕ is a scalar parameter, a_i is a prior weight and $\nu(.)$ is a variance function. The responses y_i can be modelled using generalized linear mixed models (GLMM), in which there is a monotone and differentiable link function $g(.)$ such that $g(\mu_i) = \eta_i = x_i'\beta + z_i'b$, i.e., $g(.)$ links the conditional expectation μ_i to the linear predictor η_i. In matrix form, the GLMM can be written into

$$g(\mu) = \eta = X\beta + Zb \qquad (2)$$

where μ, $g(\mu)$ and η are vectors having components μ_i, $g(\mu_i)$ and η_i $(i = 1, 2..., n)$, respectively, while the design matrices X and Z have rows x_i' and z_i', respectively.

The random effects b are usually assumed to have some distribution F with mean zero and covariance matrix $\Sigma(\theta)$, i.e., $b \sim F(0, \Sigma(\theta))$, where θ is an $(m \times 1)$ vector of unknown variance components. The magnitude of θ can be used to measure the degree of overdispersion and correlation, e.g., arising in longitudinal studies. The distribution F may assume to be Normal, for instance, see Breslow and Clayton (1993).

For the GLMM the integrated quasi-likelihood of (β, θ) thus takes the form

$$L(\beta, \theta) = \exp\{\ell(\beta, \theta)\} = \int \exp\{\sum_{i=1}^{n} \ell_i(\beta, \theta)\} dF(b; \theta) \qquad (3)$$

where

$$\ell_i(\beta, \theta) \propto \int_{y_i}^{\mu_i} \frac{a_i(y_i - u)}{\phi \nu(u)} du \qquad (4)$$

defines the conditional log quasi-likelihood of β given b. Accordingly, the maximum likelihood estimates (MLE) $(\hat{\beta}, \hat{\theta})$ that maximize $L(\beta, \theta)$ in (3) are rather difficult to obtain because $L(\beta, \theta)$ may involve analytically intractable integrals. In the literature Laplace approximation and MCMC techniques were used to locate the estimates, see, e.g., Breslow and Clayton (1993) and Karim and Zeger (1992).

2   Quasi-Monte Carlo Integration

In this paper we propose to use Quasi-Monte Carlo (QMC) approach to approximate the integrated quasi-likelihood $L(\beta, \theta)$ in (3). To gain insight into the QMC integration, let us first look at the classical Monte Carlo (MC) approximation. Suppose $f(\cdot)$ is an integrable function on the q-dimensional unit cube $C^q = [0, 1)^q$. Consider the integral

$$I(f) = \int_{C^q} f(x) dx \qquad (5)$$

In the MC integration a random sample $\mathcal{P}_K = \{x_k : 1 \le k \le K\}$ is drawn from the uniform distribution on C^q and the integral in (5) is then approximated by

$$\hat{I}_K(f, \mathcal{P}_K) = \frac{1}{K} \sum_{k=1}^{K} f(x_k) \qquad (6)$$

By the strong law of large number the estimate $\hat{I}_K(f, \mathcal{P}_K)$ converges to $I(f)$ with probability one as $K \to \infty$. Moreover the central limit theorem guarantees that $\hat{I}_K(f, \mathcal{P}_K)$ is asymptotically normally distributed when the sample size K is large enough. The convergence rate for the MC integration has an order $O(K^{-1/2})$, regardless of the dimension q. However, the convergence is in probability, implying the MC may behave well on average

but a particular random sample may lead to a bad approximation. We may apply multiple draws for random samples and then take the average to be the final approximation but computation load may increase dramatically. The QMC approach aims to improve the MC approximation in terms of convergence rate and computation load. The key idea is to choose integration nodes that are scattered on C^q uniformly. The reason behind this is due to the Koksma-Hlawka inequality:

$$|I(f) - \hat{I}_K(f, \mathcal{P}_K)| \leq V(f)D(\mathcal{P}_K) \tag{7}$$

where $V(f)$ is a bounded total variation of f over C^q in the sense of Hardy and Krause (Fang and Wang, 1994). $D(\mathcal{P}_K)$ is a measure of evenness of spread for the set $\mathcal{P}_K$, defined by

$$D(\mathcal{P}_K) = \sup_{x \in C^q} |U_K(x) - U(x)| \tag{8}$$

where $U(x)$ is the uniform distribution on C^q and $U_K(x)$ is the empirical distribution of $\mathcal{P}_K$. $D(\mathcal{P}_K)$ is called discrepancy of the point set $\mathcal{P}_K$. The inequality (7) implies that the absolute error of integration approximation is bounded by $D(\mathcal{P}_K)$ since $V(f)$ is a constant as long as $f(.)$ is given. The points with the smallest discrepancy are thus the best integration nodes in this sense. It can be shown that the smallest discrepancy has the order $O((\log K)^{q-1}/K)$ (Fang and Wang, 1994). Accordingly, when q is large the QMC integration has a faster convergence rate than the MC approximation. Unlike the MC approach, on the other hand, the QMC integration nodes are deterministic so that multiple draws are not necessary. Regarding construction of QMC integration nodes, one can refer to Fang and Wang (1994).

For illustration, in Figure 1 below we give 2D-plots of a MC random sample with size 100 and a QMC point set with size 55. The discrepancy values are also given underneath the plots.

(a) $D_{\mathrm{mc}} = 0.13$ (b) $D_{\mathrm{qmc}} = 0.04$
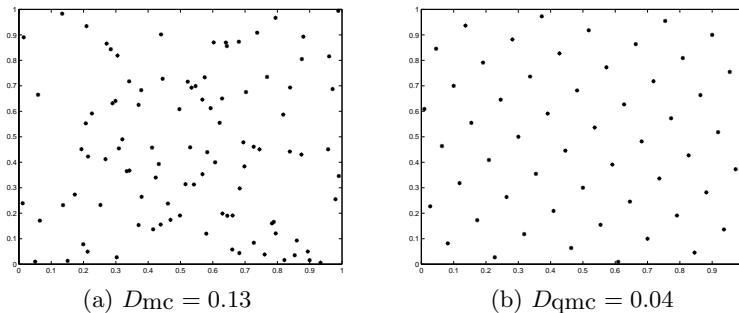
FIGURE 1. A MC random sample with size 100 (Panel (a)) and a QMC point set with size 55 (Panel (b)) over $C^2 = [0, 1)^2$

Figure 1 clearly shows that the 55 QMC nodes in Panel (b) are much better than the 100 MC random points in Panel (a) in terms of uniformity.

3   Quasi-Monte Carlo Estimation in GLMM

When applying the QMC approximation to (3), the log quasi-likelihood can be written by

$$\ell(\beta, \theta) = \log\Big(\frac{1}{K} \sum_{k=1}^{K} \exp\Big\{\sum_{i=1}^{n} \ell_i(\beta, \Sigma^{1/2} F^{-1}(b_k))\Big\}\Big) \tag{9}$$

where $\mathcal{P}_K = \{b_k : k = 1,...,K\}$ is a QMC set over C^q, $F^{-1}(.)$ is the inverse of the cdf F and $\Sigma^{1/2}$ can be taken as the Cholesky factor of Σ. Let $c_k = F^{-1}(b_k)$, $\eta_{ik} = x_i'\beta + z_i'\Sigma^{1/2}c_k$ and $\mu_{ik} = h(\eta_{ik})$ where $h(.)$ is the inverse function of $g(.)$. The MLE $\hat{\beta}$ of β then must satisfy the score equation:

$$\frac{\partial}{\partial \beta}[\ell(\beta, \theta)] = \sum_{k=1}^{K} w_k \Big[\sum_{i=1}^{n} \frac{a_i(y_i - h(\eta_{ik}))}{\phi \nu(\mu_{ik})g'(\mu_{ik})} x_i\Big] = 0 \tag{10}$$

where $g'(.)$ is the derivative of $g(.)$ and w_k has the form

$$w_k = \frac{\exp\{\sum_{i=1}^{n} \ell_i(\beta, \Sigma^{1/2}c_k)\}}{\sum_{k=1}^{K} \exp\{\sum_{i=1}^{n} \ell_i(\beta, \Sigma^{1/2}c_k)\}} \tag{11}$$

Similarly we have the score equation for the variance components θ. We further give the explicit forms for the second-derivatives of $\ell(\beta, \theta)$ and then use Newton-Raphson algorithm to calculate the MLE $(\hat{\beta}, \hat{\theta})$, which in turn gives the asymptotic variance-covariance matrix of the MLE $(\hat{\beta}, \hat{\theta})$.

4   An Example: Salamander Mating Data

The infamous salamander mating experiment involved two population of salamanders: Rough Butt (RB) and Whiteside (WS). Ten males and ten females from each population were mated in a crossd design, with six matings for each animal, resulting in 120 correlated binary observations. The experiment was repeated three times during the summer and autumn of 1986. For each experiment a logistic-Normal mixed model is used to model the correlated binary data:

$$\text{logit}\{E(y_{ij}|b_i^f, b_j^m)\} = x_{ij}'\beta + b_i^f + b_j^m \tag{12}$$

where b_i^f and b_j^m are random effects from the female and male individuals in the pair and are assumed to be independent with $b_i^f \sim N(0, \sigma_f^2)$

and $b_j^m \sim N(0, \sigma_m^2)$ $(i, j = 1, ..., 20)$. The covariate vector x_{ij} is set to be $(1, \mathrm{WS}_i^f, \mathrm{WS}_j^m, \mathrm{WS}_{ij}^{fm})$ where WS_i^f is the indicator for WS female (0=RB and 1=WS), WS_j^m for WS male (0=RB and 1=WS) and WS_{ij}^{fm} means the interaction.

The log-likelihood for each experiment is a sum of two 20-dimensional integrals which are analytically intractable (Breslow and Clayton, 1993). When pooling the three experiments data, it involves six 20-dimensional integrals. Modelling the data becomes extremely challenging. In the literature various approaches were considered, e.g., MCMC by Karim and Zeger (1992) and penalized quasi-likelihood (PQL) by Breslow and Clayton (1993).

We apply the QMC approach to modelling of the pooled data. Since the integrals are 20-dimensional, we generate QMC integration nodes on the cube $C^{20} = [0, 1)^{20}$, implemented using the first 20 prime numbers (Fang and Wang, 1994). Table 1 below gives the MLEs of the parameters, where K is the size of the QMC nodes. For comparison, we also present Karim and Zeger's (1992) Gibbs sampling and Breslow and Clayton's (1993) PQL estimates below.

Table 1. MLEs of parameters (standard errors in parentheses)

K	β_0	β_1	β_2	β_3	σ_f	σ_m	$\ell_{\max}$
10,000	0.92(.38)	-2.83(.51)	-0.58(.41)	3.57(.63)	1.11(.28)	0.98(.20)	-207.21
20,000	0.83(.37)	-2.80(.52)	-0.53(.44)	3.51(.61)	1.06(.23)	1.02(.23)	-207.70
30,000	1.28(.41)	-2.88(.54)	-0.99(.50)	3.64(.63)	1.25(.27)	1.16(.24)	-205.67
40,000	1.22(.41)	-2.83(.53)	-0.99(.49)	3.66(.63)	1.28(.28)	1.21(.26)	-206.19
50,000	1.21(.40)	-2.81(.53)	-1.03(.49)	3.70(.62)	1.25(.24)	1.24(.26)	-206.07
60,000	1.17(.39)	-2.80(.53)	-0.99(.49)	3.67(.63)	1.23(.24)	1.20(.26)	-206.41
70,000	1.21(.37)	-2.81(.53)	-0.96(.47)	3.68(.63)	1.30(.29)	1.22(.26)	-206.31
80,000	1.22(.38)	-2.86(.54)	-1.01(.49)	3.71(.64)	1.30(.29)	1.24(.26)	-206.35
90,000	1.21(.38)	-2.87(.54)	-0.99(.49)	3.69(.64)	1.28(.29)	1.22(.26)	-206.66
100,000	1.22(.39)	-2.91(.56)	-0.98(.49)	3.67(.64)	1.26(.29)	1.23(.27)	-206.83
Gibbs	1.03(.43)	-3.01(.60)	-0.69(.50)	3.74(.68)	1.22	1.17	–
PQL	0.79(.32)	-2.29(.43)	-0.54(.39)	2.82(.50)	0.85	0.79	–

Table 1 above shows that even for such high-dimensional integrals the QMC approach can do a good job by choosing an appropriate size of the QMC nodes. We also discuss hypothesis test for variance components using score test. Simulation studies to mimic the salamander data are also conducted.

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.

Fang, K. T. and Wang Y. (1994). *Number-Theoretic Methods in Statistics*, Chapman and Hall, London.

Karim, M. R. and Zeger, S. L. (1992). Generalised linear models with random effects: salamander mating revisited *Biometrics*, **48**, 631-644.

Modelling Covariance Structures in Generalized Estimating Equations for Longitudinal Data

Huajun Ye[1] and Jianxin Pan[1]

[1] Mathematics Department, Manchester University, Manchester, M13 9PL, U.K.
Email: yhuajun@maths.man.ac.uk; jpan@maths.man.ac.uk

Abstract: Generalized estimating equations (GEE) (Liang & Zeger, 1986) are commonly used for longitudinal studies. In the literature little attention was paid to the choice of "working" covariance structures in GEE. Recently Wang & Carey (2003) showed that mis-specification of "working" covariance structures may lead to a great loss of estimate efficiency. In this paper we propose a data-driven approach for modelling covariance structures in GEE for longitudinal data. Our numerical analysis shows that the GEE estimate efficiency can be improved in terms of modelling of covariance structures.

1   Generalized estimating equations

Consider a longitudinal study protocol. Let y_{ij} be the jth of m_i measurements on the ith of n subjects. Assume t_{ij} are the time at which the measurement y_{ij} are made. Denote the responses of the ith subject by $y_i = (y_{i1}, y_{i2}, ..., y_{im_i})'$ and the time points by $t_i = (t_{i1}, t_{i2}, ..., t_{im_i})'$. Suppose $E(y_i) = \mu_i$ and $Var(y_i) = \Sigma_i$ are the $(m_i \times 1)$ mean vector and $(m_i \times m_i)$ variance-covariance matrix of y_i, respectively.

The mean μ_{ij} is usually related to some covariates of interest, say x_{ij} (e.g., x_{ij} may contain t_{ij}), through a link function: $g(\mu_{ij}) = x'_{ij}\beta$. In longitudinal studies, we might be only concerned with the estimate of the parameter vector β $(p \times 1)$ regardless of the structures of Σ_i. Accordingly, certain "working" covariance structures are used to model Σ_i and then to solve the generalized estimating equations (GEE):

$$S(\beta) = \sum_{i=1}^{n} \left[\frac{\partial \mu'_i}{\partial \beta}\right] V_i^{-1/2} C_i^{-1}(\rho) V_i^{-1/2} (y_i - \mu_i) = 0 \qquad (1)$$

(Liang & Zeger, 1986) where $V_i = \text{diag}(v_{i1}^2, ..., v_{im_i}^2)$ with $v_{ij} = Var(y_{ij})$. The matrix $C_i(\rho)$ that depends on a new scalar parameter ρ mimics the

within-subject correlation, for instance, it may take compound structure or AR(1), etc. Under certain regularity conditions, it may be shown that the GEE estimates are asymptotical Normally distributed and consistent (Liang & Zeger, 1986).

2   Modelling covariance structures in GEE

Recently there is an increasing concern about the mis-specification of the "working" covariance structures. When the covariance structures are mis-specified, the efficiency of the GEE estimates $\hat{\beta}$ may be rather poor although it is consistent (Wang, 2003). Accordingly, we want to model the covariance structures together with estimating β.

Since Σ_i is positive definite, there exists a unique lower triangular matrix T_i with 1's as diagonals and a unique diagonal matrix D_i with positive diagonals such that $T_i \Sigma_i T_i^{'} = D_i$. This modified Cholesky decomposition has a clear statistical interpretation: the below-diagonals of T_i are the negatives of the autoregressive coefficients, ϕ_{ijk}, in the autoregression model

$$\hat{y}_{ij} = \mu_{ij} + \sum_{k=1}^{j-1} \phi_{ijk}(y_{ik} - \mu_{ik}) \tag{2}$$

and the diagonals of D_i are the innovation variances $\sigma_{ij}^2 = Var(\varepsilon_{ij})$ where $\varepsilon_{ij} = y_{ij} - \hat{y}_{ij}$ $(1 \leq j \leq m_i; 1 \leq i \leq n)$.

In a spirit of Pourahmadi (1999), we propose three generalized regression models to model the mean, autoregressive coefficients and innovation variances:

$$g(\mu_{ij}) = x'_{ij}\beta, \quad \phi_{ijk} = z'_{ijk}\gamma \quad \text{and} \quad \log \sigma_{ij}^2 = z'_{ij}\lambda \tag{3}$$

where the covariates x_{ij}, z_{ijk} and z_{ij} are $(p \times 1)$, $(q \times 1)$ and $(d \times 1)$ vectors, respectively, and β, γ and λ are the associated parameters. The link function $g(.)$ is assumed to be monotone and differentiable (McCullagh & Nelder, 1989).

In order to estimate β, γ and λ in (3), we propose to solve the three generalized estimating equations as follows:

$$
\begin{aligned}
S_1(\beta) &= \sum_{i=1}^{n} \Big[\frac{\partial \mu'_i}{\partial \beta}\Big] \Sigma_i^{-1}(y_i - \mu_i) \\
S_2(\gamma) &= \sum_{i=1}^{n} \Big[\frac{\partial \hat{r}'_i}{\partial \gamma}\Big] D_i^{-1}(r_i - \hat{r}_i) \\
S_3(\lambda) &= \sum_{i=1}^{n} \Big[\frac{\partial \sigma_i^{2'}}{\partial \lambda}\Big] W_i^{-1}(\epsilon_i^2 - \sigma_i^2)
\end{aligned}
\tag{4}
$$

where in the second equation r_i and $\hat{r}_i$ are $(m_i \times 1)$ vectors with the jth components $r_{ij} = y_{ij} - \mu_{ij}$ and $\hat{r}_{ij} = E(r_{ij}|r_{i1}, ..., r_{i(j-1)}) = \sum_{k=1}^{j-1} \phi_{ijk}r_{ik}$

$(j = 1, ..., m_i)$, respectively. It can be shown that $D_i = \text{diag}(\sigma_{i1}^2, ..., \sigma_{im_i}^2)$ are in fact the covariance matrix of $r_i - \hat{r}_i$. In the third equation ϵ_i^2 and σ_i^2 are $(m_i \times 1)$ vectors with the jth components ϵ_{ij}^2 and σ_{ij}^2 $(j = 1, ..., m_i)$, respectively, where $\epsilon_{ij} = y_{ij} - \hat{y}_{ij}$ and $\hat{y}_{ij}$ are given in (2). Obviously, we have the fact $E(\epsilon_i^2) = \sigma_i^2$. In addition, W_i is the covariance matrix of ϵ_i^2, i.e., $W_i = Var(\epsilon_i^2)$.

When data are Normally distributed, we can show $W_i = 2\text{diag}(\sigma_{i1}^4, ..., \sigma_{im_i}^4)$ so that (4) reduces to Pourahmadi's (1999) score equations in this special case. In general, however, W_i may not be diagonal and should be estimated together with other parameters. In the spirit of traditional GEE modelling for the mean, we specify a sandwich "working" structure to W_i, say $W_i = A_i^{1/2} R_i(\rho) A_i^{1/2}$ where $A_i = 2\text{diag}(\sigma_{i1}^4, ..., \sigma_{im_i}^4)$ and $R_i(\rho)$ mimics the correlations between ϵ_{ij}^2 and ϵ_{ik}^2 $(i \neq k)$ in terms of a new parameter ρ. Typical examples include compound symmetry and AR(1).

We propose an algorithm to iteratively calculate the solutions $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\lambda}$ to (4), which are termed GEE estimates for β, γ and λ. Under certain regularity conditions, we showed that $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\lambda}$ are consistent and asymptotically Normal. We also consider hypothesis tests regarding β, γ and λ based on score test principles.

3   Numerical analysis

We analyze Kenward's cattle data in which 60 animals were assigned randomly to two treatment groups A and B. Half animals received treatment A and another half received treatment B. The cattles were weighted 11 times over 133-day period at 0, 14, 28, 42, 56, 70, 84, 98, 112, 126 and 133 in days and the objective was to study treatment effects on intestinal parasites.
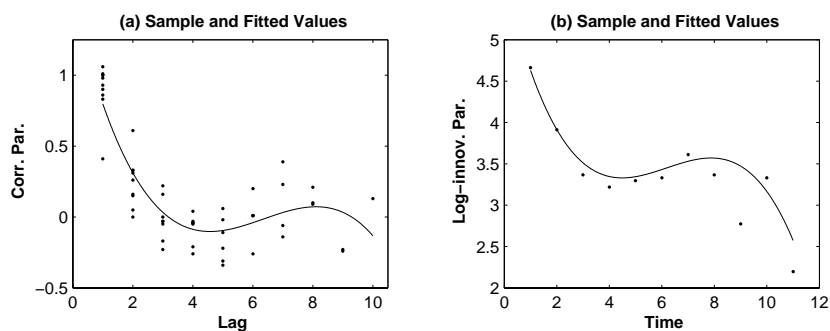
Figure 1. The sample regressogram and the proposed GEE fitted curves

For illustration we only model the Treatment A data here. Following Pourahmadi's (1999) protocol we use a saturated model for the mean and choose two cubic polynomials of time/lag to model the innovation variances and autoregressive coefficients. In the modelling we choose the function $g(.)$

as identity link and specify the "work" correlation structure for $R_i(\rho)$ as compound symmetry and AR(1). For both "working" structures we testify the GEE estimates by choosing different values of ρ, e.g., at $\rho = 0.2, 0.5$ and 0.8. We find that the parameter ρ, measuring the correlation between ϵ_{ij}^2 and ϵ_{ik}^2, affects very little the estimates of γ and λ. This implies that the GEE estimates are robust against the possible mis-specification of the structure of $R_i(\rho)$. This point has been also confirmed from our simulation studies below. In Figure 1 above we plot the sample autoregressive coefficients and sample log-innovation variance (dot points) and also display the GEE fitted curves with $R_i(\rho)$ being AR(1) where $\rho = 0.5$, which clearly shows that the proposed GEE approach fits the data well.

In order to measure the efficiency of estimates for fixed effects, we propose to use a cubic polynomial of time rather than the saturated model to model the trajectory of mean (Pan & MacKenzie, 2003). Table 1 below gives the comparison of the proposed approach with the conventional GEE estimates in terms of relative efficiency of the fixed effects β_i ($i = 1, ..., 4$). The relative efficiency of β_i is defined as the ratio of variance of the conventional GEE estimate $\hat{\beta}_i^C$ resulted from (1) to that of the covariance modelling GEE estimate $\hat{\beta}_i^M$ obtained by solving (4), i.e., $e(\hat{\beta}_i) = Var(\hat{\beta}_i^C)/Var(\hat{\beta}_i^M)$. Both compound symmetry (CS) and AR(1) are used to be "working" covariance structures and the correlation parameter ρ is set to be the same in the conventional and the new GEE estimation procedures.

Table 1. Relative efficiency for fixed effects

	CS			AR(1)		
ρ	0.2	0.5	0.8	0.2	0.5	0.8
$e(\beta_1)$	1.42	1.29	1.11	1.24	1.42	1.47
$e(\beta_2)$	1.17	1.19	1.23	1.21	1.14	1.14
$e(\beta_3)$	1.37	1.33	1.32	1.37	1.26	1.26
$e(\beta_4)$	1.21	1.20	1.17	2.12	2.07	1.78

Table 1 above shows that the efficiency of the conventional GEE estimates can be improved in terms of covariance modelling strategy. In some cases the variance of $\hat{\beta}_i^C$ may be twice of variance of $\hat{\beta}_i^M$.

4   Simulation Study

We conduct a simulation study for Normal and Normal mixture. For Normal, Table 2 below gives the comparison of the proposed approach to the conventional GEE estimates in terms of averaged relative efficiency of the fixed effects β_i ($i = 1, ..., 4$), where we generate $30 \times 10,000$ random numbers from the Normal distribution with the mean vector μ_i and variance matrix Σ_i obtained from the Cattle data. For Normal mixture, we choose the distribution $F = \pi N(\mu_i + \delta, \Sigma_i) + (1 - \pi)N(\mu_i, \Sigma_i)$ where mean vector

μ_i and variance matrix Σ_i are the same as above. We generate $30 \times 10,000$ random numbers from normal mixture with $\pi = 0.5$ and $\delta = \mu_i/5$, Table 3 below gives the comparison of averaged relative efficiency between the proposed approach and the conventional GEE estimates.

Table 2. Averaged relative efficiency for Normal distribution

	CS			AR(1)		
ρ	0.2	0.5	0.8	0.2	0.5	0.8
$e(\beta_1)$	1.38	1.39	1.39	1.19	1.35	1.41
$e(\beta_2)$	1.15	1.15	1.16	1.18	1.11	1.12
$e(\beta_3)$	1.34	1.33	1.34	1.32	1.20	1.21
$e(\beta_4)$	1.18	1.19	1.18	2.05	2.01	1.72

Table 3. Averaged relative efficiency for Normal mixture distribution

	CS			AR(1)		
ρ	0.2	0.5	0.8	0.2	0.5	0.8
$e(\beta_1)$	1.15	1.12	1.04	1.09	1.15	1.16
$e(\beta_2)$	1.10	1.13	1.23	1.06	1.05	1.06
$e(\beta_3)$	1.40	1.39	1.36	1.34	1.24	1.26
$e(\beta_4)$	1.34	1.31	1.23	2.67	2.43	2.07

Table 2 and Table 3 above show that covariance modelling strategy improves the efficiency of the conventional GEE estimates. In some cases, the improvement is very significant, implying that mis-specification of the "working" covariance structure in GEE may lead to inefficient estimates of fixed effects. Accordingly, correctly modelling covariance structure plays an important role in GEE procedure.

Acknowlegement:

References

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

McCullagh, P. and Nelders, J.A. (1989). *Generalized Linear Models* (second edition), Chapman and Hall.

Pan, J. and MacKenzie, G. (2003). Model selection for joint mean-covariance structures in longitudinal studies. *Biometrika*, **90**, 239-244.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, **86**, 677-90.

You-Gan Wang and Vincent Carey (2003). Working correlation structure mis-specification, estimation and covariate design: Implications for generalized estimating equations performance. *Biometrika* **90**, 29-41.

Count distributions with mixed Poisson random effects

Pedro Puig[1] and Jordi Valero[2]

[1] Universitat Autònoma de Barcelona, Spain, ppuig@mat.uab.es
[2] Escola Superior d'Agricultura de Barcelona UPC, Spain, jordi.valero@upc.es

1   Introduction

Consider independent count observations y_i with covariate explicative vectors x_i, $i = 1, ..., n$. Poisson regression models assume that $y_i \sim \text{Poisson}(\mu_i)$, where $\mu_i = \mu_i(x_i, \beta)$ and β is a k-dimensional vector of unknown parameters.

When overdispersion occurs or repeated measures are done, mixed Poisson models are frequently used in the form, $y_i \sim \text{Poisson}(\mu_i \varepsilon_i)$, where ε_i are iid positive random variables, such that $E(\varepsilon_i) = 1$ and $\text{var}(\varepsilon_i) = \sigma^2$. This leads to the second order variance function $\text{var}(y_i; x_i) = \mu_i + \sigma^2 \mu_i^2$ (Collings and Margolin, 1985). For instance, it is well known that if the ε_i's are assumed to have a gamma distribution, then y_i follows a negative binomial distribution. If the ε_i's follow an inverse gaussian distribution, then the distribution of y_i is known as Poisson-Inverse Gaussian. A good reference about mixed Poisson models can be found in Lawless (1987).

2   The models

In (Puig and Valero, 2004) the following concept of additivity is introduced:

Definition. *Given a parametric model we shall say that it is "partially closed under addition" if for each random variable X belonging to this model the sum of any number of independent copies of X also belongs to this parametric model.*

These kind of models can arise naturally in many practical situations. For instance, many data sets come from counts on independent quadrats or sub-areas of an experimental region. If we consider that some parametric model is valid to describe these counts, it is reasonable to hope that the same model is valid to describe the counts made by grouping two or more quadrats.

Our aim is to find all the mixed Poisson models that are partially closed under addition and have good properties about the estimation of the population mean. The following theorem is a direct consequence of the results of Puig and Valero (2004) and Puig (2003):

Theorem. *Let Y be a mixed Poisson model, $Y = Poisson(\mu\varepsilon)$, such that $E(\varepsilon) = 1$, $var(\varepsilon) = \sigma^2$ and $E(\varepsilon^3) < \infty$, with a pgf continuous in μ and twice differentiable with continuity in σ^2. Assume that Y is partially closed under addition and the MLE of μ is the sample mean. Then its probability generating function (pgf) can be expressed as,*

$$g(t; \mu, \sigma^2, \beta) = e^{\frac{1-\beta}{\beta\sigma^2}\left[1-(1-\frac{\mu\sigma^2}{1-\beta}(t-1))^\beta\right]} \tag{1}$$

The domain of β is $\beta \leq 1$.

For $\beta = -1$ and $\beta = 1/2$ we obtain directly the pgf of the Polya-Aeppli and Poisson-Inverse Gaussian distribution. Calculating the limit when β tends to $-\infty$ and 0, we get respectively the pgf of the Neyman A and the Negative Binomial distribution. In general for $\beta < 1$ this family is known as Poisson-Tweedie distribution or Power Variance Mixture model (Hougaard et al., 1997). The Tweedie densities have not in general a simple expression. However, when they act as a mixing distribution, the resulting mixed Poisson distribution has a simple pgf.

In the next section we shall show how we can implement some simple correlational structures between count data using the Tweedie models.

3   Mixed Poisson with random effects

3.1   Paired count data

Given the paired count data y_{ij} $i = 1, 2$, $j = 1, ..., n$, we assume that its distribution is of the form $Poisson(\mu_i\epsilon_i^*)$ where $\epsilon_1^* = \lambda_1\epsilon_1 + (1 - \lambda_1)\epsilon_0$, $\epsilon_2^* = \lambda_2\epsilon_2 + (1 - \lambda_2)\epsilon_0$, and $\lambda_i \in [0, 1]$ are two new parameters. The random variables ϵ_1, ϵ_2 and ϵ_0 are independent members of the Tweedie family, with expectation equal to 1, variances σ_1^2, σ_2^2 and σ_0^2, and parameters β_1, β_2 and β_0 respectively. Notice that ϵ_0 has the same value for the two members of the same couple. It can be interpreted as the perturbation due to the random effect "couple". Consequently, from (1) and direct calculations, the joint log-pgf for the paired observations (y_{1j}, y_{2j}) remains,

$$
\begin{aligned}
\log(g(t_1, t_2)) = & \frac{1-\beta_1}{\beta_1\sigma_1^2}\left[1 - (1 - \frac{\mu_1\lambda_1\sigma_1^2}{1-\beta_1}(t_1 - 1))^{\beta_1}\right] \\
& + \frac{1-\beta_2}{\beta_2\sigma_2^2}\left[1 - (1 - \frac{\mu_2\lambda_2\sigma_2^2}{1-\beta_2}(t_2 - 1))^{\beta_2}\right] \\
& + \frac{1-\beta_0}{\beta_0\sigma_0^2}\left[1 - (1 - \frac{\mu_1(1-\lambda_1)(t_1-1)+\mu_2(1-\lambda_2)(t_2-1)}{1-\beta_0}\sigma_0^2)^{\beta_0}\right].
\end{aligned}
\tag{2}
$$

From (2) it is immediate to find that,

$Var(y_{ij}) = \mu_i + \mu_i^2 \left(\lambda_i^2 \sigma_i^2 + (1 - \lambda_i)^2 \sigma_0^2 \right)$, and also the covariance

$cov(y_{1j}, y_{2j}) = \mu_1 (1 - \lambda_1) \mu_2 (1 - \lambda_2) \sigma_0^2$.

A naive approach to the paired count data problem could suggest to consider only the perturbation due to the "couple" random effect, that is, to fix $\lambda_1 = \lambda_2 = 0$. However, in this situation, the correlation coefficient of the paired observations is absolutely determined by the dispersion indexes δ_i of the marginals, that is, $r(y_{1j}, y_{2j})^2 = (\delta_1 - 1)(\delta_2 - 1)/(\delta_1 \delta_2)$. Consequently this naive model is not very flexible in practice.

Example 1: In an experiment of Agriculture we count the feasible seeds of *Digitaria sanguinalis* according to a minimum tillage (TS) or no tillage at all (SD) of the soil. We have 72 blocks, and we have counted a sample of TS and SD for each block. The results of the experiment can be summarized as follows:

Tillage	Mean	Variance	disp. index	**Corr. coeff.**
TS	2.778	28.288	10.184	0.364
SD	0.417	1.092	2.620	

Notice that, if the naive model previously commented was adequate, a correlation coefficient about 0.75 can be predicted, from the empirical dispersion indexes shown above. However the empirical correlation coefficient is about 0.36.

In order to analize the data set we are going to use the full model with the restriction $\beta_0 = \beta_1 = \beta_2 = \beta$. The corresponding probabilities can be computed from (2), and the program made in R that we have performed gives the maximum likelihood estimators:

$\log(L)$	$\hat{\mu}_{TS}$	$\hat{\mu}_{SD}$	$\hat{\lambda}_{TS}$	$\hat{\lambda}_{SD}$	$\hat{\sigma}_{TS}^2$	$\hat{\sigma}_{SD}^2$	$\hat{\sigma}_0^2$	$\hat{\beta}$
-202.3	2.778	.417	.802	.464	4.98	$\simeq 0$	26.54	0.48

¿From here, the estimated variances and dispersion indexes of the marginals are $\hat{V}_{TS} = 35.502$, $\hat{V}_{SD} = 1.742$, $\hat{\delta}_{TS} = 12.781$, $\hat{\delta}_{SD} = 4.180$, and the estimated correlation coefficient is now $\hat{r} = 0.415$. Notice that these estimated values are similar to the empirical values shown above. The estimated value of β is close to $1/2$, that is, the Tweedie model of the ε's is close to the Inverse Gaussian distribution.

Likelihood ratio tests can be performed in order to check if the model can be simplified and to compare the means of the counts of feasible seeds according the kind of tillage:

H_0			df	χ^2	p value
$\lambda_{TS} = \lambda_{SD}$	$\sigma_{TS}^2 = \sigma_{SD}^2$	$\mu_{TS} = \mu_{SD}$	3	34.945	< 0.001
$\lambda_{TS} = \lambda_{SD}$	$\sigma_{TS}^2 = \sigma_{SD}^2$		2	6.529	0.038
		$\mu_{TS} = \mu_{SD}$	1	28.415	< 0.001

Consequently the tillage takes effect on the abundance of feasible seeds. It is also interesting to test the significance of the "couple" random effect,

that is, to consider the null hypothesis $H_0 : \lambda_{TS} = \lambda_{SD} = 0$. The resulting likelihood ratio test statistic is 20.082 with a p-value $p < 0.001$. It is important to remark that, under the null hypothesis, the likelihood ratio test statistic does not have an asymptotic χ^2 distribution as expected, because $\lambda_{TS} = \lambda_{SD} = 0$ belongs to the boundary of the domain of parameters (see Self and Liang, 1987).

3.2   Implementing a random effect

Now we study a simple generalization of the situation presented in the preceding section. We consider the case y_{ij} $i = 1, ..., k$ $j = 1, ..., n$, where its distribution follows a Poisson$(\mu_i \epsilon_i^*)$ with $\epsilon_i^* = \lambda_i \epsilon_i + (1 - \lambda_i) \epsilon_0$, $\lambda_i \in [0, 1]$. The random variables ϵ_i and ϵ_0 are independent members of the Tweedie family, with expectation equal to 1, variances σ_i^2 and σ_0^2, and parameters β_i and β_0 respectively. Now ϵ_0 can be understood as the perturbation due to the random effect of the group or block. Direct calculations give the log-pgf:

$$
\begin{aligned}
\log(g(t_1, ..., t_k)) = &\sum_{i=1}^{k} \frac{1-\beta_i}{\beta_i \sigma_i^2} \left[1 - (1 - \frac{\mu_i \lambda_i \sigma_i^2}{1-\beta_i} (t_i - 1))^{\beta_i} \right] \\
&+ \frac{1-\beta_0}{\beta_0 \sigma_0^2} \left[1 - (1 - \frac{\sigma_0^2}{1-\beta_0} \sum_{i=1}^{k} \{ \mu_i (1 - \lambda_i) (t_i - 1) \})^{\beta_0} \right].
\end{aligned}
\tag{3}
$$

This model, in the most general situation, has $4k + 2$ parameters. Some of them, like β_i or σ_i^2, can be assumed to be equal in order to simplify the model. The variances have the same expression like in the case of paired data, and the covariances are $cov(y_{rj}, y_{sj}) = \mu_r (1 - \lambda_r) \mu_s (1 - \lambda_s) \sigma_0^2$.

Example 2: Here the aim of the experiment is to study the relation between the abundance of three kind of feasible seeds *Polygonum aviculare*, *Portulaca oleracea* and *Diplotaxis erucoides*, under a minimum tillage of the soil. The sample comes from 72 points where the three kind of seeds have been counted. The results of the experiment can be summarized as follows:

Seed	Mean	Variance	disp. index	pair	Covariance	Corr.
Poly	2.236	3.817	1.707	$Pol - Por$	0.594	0.235
Port	0.639	1.671	2.615	$Pol - Dip$	0.935	0.234
Dipl	0.861	4.178	4.851	$Por - Dip$	0.189	0.071

We have fitted the data set considering the model where the β's are equal. The maximum of the log-likelihood function is $\log(L) = -295.805$ and the values of the estimators are as follows:

$\hat{\mu}_{Pol}$	$\hat{\mu}_{Por}$	$\hat{\mu}_{Dip}$	$\hat{\lambda}_{Pol}$	$\hat{\lambda}_{Por}$	$\hat{\lambda}_{Dip}$	$\hat{\sigma}_{Pol}^2$	$\hat{\sigma}_{Por}^2$	$\hat{\sigma}_{Dip}^2$	$\hat{\sigma}_0^2$	β
2.236	0.639	0.861	0.91	0.88	0.32	0.36	3.66	0.19	8.32	-1.6

Using these values we can also estimate the variances, correlation coefficients, etc. The results are the following:

Seed	*Mean*	*Variance*	*Disp. index*	**pair**	*Covariance*	*Corr.*
Poly	2.236	4.094	1.831	$Pol - Por$	0.132	0.048
Port	0.639	1.848	2.892	$Pol - Dip$	1.021	0.261
Dipl	0.861	3.742	4.345	$Por - Dip$	0.371	0.141

The resemblance with the empirical results is notorious. However the predicted correlation of $Pol - Por$ is lower than the empirical value. To check if the random effect is significant we have to consider the null hypothesis $H_0 : \lambda_{Pol} = \lambda_{Por} = \lambda_{Dip} = 0$. The resulting likelihood ratio test statistic is 12.3308 with a p-value $p < 0.001$. Consequently, the abundance of any kind of the three studied seeds is correlated with the others.

References

Collins, B.J. and Margolin, B.H. (1985). Testing goodness of fit for the Poisson assumption when observations are not identically distributed. *J. Amer. Statist. Assoc.*, **80**, 411–418.

Hougaard, P., Ting Lee, M.L. and Whitmore G.A. (1997). Analysis of Overdispersed Count Data by Mixtures of Poisson Variables and Poisson Processes. *Biometrics*, **53**, 1225–1238.

Lawless, J.F. (1987). Negative Binomial and mixed Poisson regression. *Canadian Journal of Statistics*, **15**, 209–225.

Puig, P. (2003). Characterizing additively closed discrete models by a property of their MLEs, with an application to generalized Hermite distributions. *J. Amer. Statist. Assoc.*, **98**, 687–692.

Puig, P. and Valero, J. (2004). Count data distributions: some characterizations with applications. Submitted to *J. Amer. Statist. Assoc.*

Self, S. G. and Liang, K.Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.*, **82**, no. 398, 605–610.

Improving the Relevance Vector Machine under Covariate Measurement Error

David Rummel

[1] Department für Statistik, Universität München, 80539 München, Germany

Abstract: Covariate measurement error is an identified problem in statistical analysis applying parametric and nonparametric regression models. We investigate this problem for a very recent and promising smoothing approach coming from the area of machine learning, the relevance vector machine (RVM), developed by Tipping (2000). Two standard correction methods for measurement error, regression calibration (Carroll et al. (1995)) and the so-called SIMEX method (Carroll et al. (1999)), are discussed and applied to the RVM. Finally, we present a short simulation study on both methods that indicates improvements of the RVM regression in terms of bias and mean squared error.

Keywords: Nonparametric regression, automatic relevance determination, covariate measurement error, SIMEX, regression calibration.

1 Introduction

Nonparametric regression has been widely established in statistical analysis and of particular interest are simple models that allow for highly flexible data approximation. We focus here on nonparametric regression with the relevance vector machine, as introduced by Tipping (2000). Covariates surveyed under measurement error is a popular problem in the area of medicine and epidemiology, where e.g. the exposure to a certain radiation or nutrition has to be recorded. Especially in the case of nonparametric regression this covariate measurement error problem has not received much attention, yet.

The first section provides insight into the model specification of the RVM, while the second presents some theoretic background on measurement error correction. Finally we present the results of a short simulation study on correcting for error applying the SIMEX method and regression calibration.

2 Nonparametric regression using the RVM

Generally we are given data of the form $\{(\mathbf{x}_i, t_i)\}_{i=1}^{N} \in R^D \times R$ including a D-dimensional vector of covariates $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD})$ and a scalar target t_i for each observation i. We note, that covariate measurement error is not an issue in this section.

2.1   The RVM model setup

It is assumed for the RVM, that the dependency of the targets on the covariates can be represented by a sum of basis functions $\phi_j(\mathbf{x})$, individually weighted by a related parameter w_j and an intercept w_0. Since the targets are generally assumed to consist of a structural and a random part, we have here:

$$t_i = \sum_{j=1}^{N} w_j \phi_j(\mathbf{x}_i) + w_0 + \epsilon_i, \quad i = 1, \ldots, N, \tag{1}$$

where the errors are assumed to be i.i.d. normally distributed $p(\boldsymbol{\epsilon}) = \prod_{i=1}^{N} \mathcal{N}(\epsilon_i|0, \sigma^2)$. By specifying (1) we allow *every* observation to have an individual impact on the structural part. To construct a model that is able to infer automatically which basis are most relevant for the regression, Tipping (2000) follows an approach of MacKay (1994), termed *automatic relevance determination*. The preference for a sparse model with only few weights being nonzero is encoded by placing a Gaussian prior over every weight, centered on zero with an individual variance parameter:

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{j=0}^{N} \mathcal{N}(w_j|0, \alpha_j^{-1}), \quad \mathbf{w} = (w_0, w_1, \ldots, w_K)^{\mathrm{T}} \tag{2}$$

To put the RVM into a fully Bayesian framework, Tipping (2000) specifies Gamma (hyper-) priors for the inverse variance parameters $p(\boldsymbol{\alpha}) = \prod_{j=0}^{N} Gamma(\alpha_j|a, b)$ and $p(\beta) = Gamma(\beta|c, d)$, setting the corresponding parameters $a = b = c = d = 0$, which is equivalent to specifying uniform distributions for $\boldsymbol{\alpha}$ and β on a logarithmic scale.

2.2   Inference

Estimation of the unknown parameters $\mathbf{w}, \boldsymbol{\alpha}$ and β in a Bayesian framework is done via the posterior distribution of these parameters:

$$p(\mathbf{w}, \boldsymbol{\alpha}, \beta|\mathbf{t}) = p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}, \beta)p(\boldsymbol{\alpha}, \beta|\mathbf{t}), \tag{3}$$

with $p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}, \beta)$ being Gaussian, see Tipping (2001) for details. Since the posterior of the hyperparameters $\boldsymbol{\alpha}, \beta$ can not be stated, Tipping (2001) suggests to find the modus of $p(\boldsymbol{\alpha}, \beta|\mathbf{t})$. Since $p(\boldsymbol{\alpha})$ and $p(\beta)$ are uniform (over a logarithmic scale), we just maximize the marginal likelihood $p(\mathbf{t}|\boldsymbol{\alpha}, \beta)$. In similar Bayesian models, this maximizing method is referred to as type-II maximum likelihood method.

Inference on the unknown parameters $\mathbf{w}, \boldsymbol{\alpha}$ and β yields a final estimation for $\widehat{f}(\mathbf{x}) = \sum_{j=1}^{N} \widehat{w}_j \phi_j(\mathbf{x}_i) + w_0$, where only very few weights $w \neq 0$ remain in the model. The data points related to these bases are then called *relevant vectors* in deference to that method.

Tipping (2001) compares the performance of the relevance vector machine to the support vector machine, another popular method in the machine learning area, and states good results for benchmark data sets.
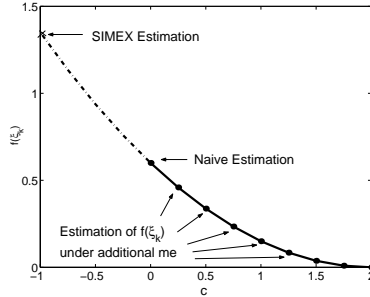
FIGURE 1. By inflating the variance for artificially generated error by $c \times \sigma_\delta^2$, the effect of additional error on the estimation $\widehat{f}(\xi_k)$ can be studied. For $c = 0$ we use the original data in the analysis. The curve can be extrapolated to the case of zero measurement error by using quadratic regression.

3   Covariate measurement error

From now on we include covariate measurement error into our considerations. That is, we distinguish between the true (but latent) covariate ξ and the observable version X under measurement error. Statistical analysis ignoring such inherent error is referred to as 'naive analysis'.

3.1   The classical error model

To take measurement error into account for statistical analysis, we need to construct a model, relating the true covariate to the observable covariate. Assume that there is a true covariate ξ but our device allows measurement merely under inclusion of a random error. We model that type of error as:

$$X = \xi + \delta, \quad (\delta, \xi) \sim \text{indep.}, \quad E(\delta) = 0, \tag{4}$$

which is frequently extended to $\delta \sim \mathcal{N}(0, \sigma_\delta^2)$ and $\xi \sim \mathcal{N}(\mu_\xi, \sigma_\xi^2)$.
There are two standard approaches to error correction, which we will discuss for the RVM successively: Carroll et al. (1999) present one adoption of the SIMEX approach for nonparametric regression and Carroll et al. (1995) describe regression calibration.

3.2   Error correction using SIMEX

The effect of covariate measurement error on the estimation function is studied in a simulation study and afterwards an extrapolation on the error-free case is performed.
For the classical error model (4), we generate random errors $\delta_i^* \sim N(0, \sigma_{\delta^*}^2)$, add these to the sampled $\mathbf{x}_i$'s and perform a standard RVM analysis using these 'new' data under additional error. Varying the error variances $\sigma_{\delta^*}^2 \equiv c \cdot \sigma_\delta^2$ allows us to study its effect on the prediction $\widehat{f}(\xi_k)$. Figure 1

TABLE 1. MSE for naive, SIMEX and regression calibration correction.

σ_δ^2	naive	SIMEX	reg. calib.
$\sigma_\delta^2 = 0.25$	0.0044	0.0055	0.0038
$\sigma_\delta^2 = 1$	0.0106	0.0103	0.0093
$\sigma_\delta^2 = 4$	0.0394	0.0256	0.0199
$\sigma_\delta^2 = 9$	0.0708	0.0585	0.0288

illustrates the increasing attenuation of $\widehat{f}(\xi_k)$ with increasing variance of the additional error. Finally we extrapolate on the case of zero measurement error. The error variance σ_δ^2 needs to be known or estimated, e.g. from validation or replication data.

3.3   Error correction using regression calibration

Carroll et al. (1995) describe the principle of regression calibration. From the model structure of the RVM (1) it follows for the case of an error prone covariate, that the mean of T given X can be written in two ways:

$$E(T|X) = \begin{cases} \sum_{j=1}^{N} w_j \, E(\phi_j(\xi)|X) + w_0 & (a) \\ \sum_{j=1}^{N} w_j^* \phi_j(X) + w_0^* & (b) \end{cases} \tag{5}$$

We note, that plugging $E(\phi_j(\xi)|X)$ (instead of $\phi_j(X)$) into the model, maintains the original weights w_j in (a), whereas usage of the error prone variable X corresponds to biased weights $w_j^* \neq w_j$ in (b). Under a parametric model for ξ given X, the conditional expectation in (5, a) is easily calculated. Replacing $\phi_j(\boldsymbol{\xi}_i)$ by $E(\phi_j(\boldsymbol{\xi}_i)|X)$ in the optimization algorithm of the RVM leads to the estimation of the original model parameters $\mathbf{w}$. We note that the error variance σ_δ^2 again has to be estimated or known.

3.4   Simulation results

We extended a RVM program code by Michael Tipping, which can be found at `http://research.microsoft.com/mlp/RVM/relevance.htm` to both the SIMEX and regression calibration case.

To check the performance of both methods, we ran 200 simulations with the following setup: 201 samples were generated from the true function $f(\xi) = \sin(\xi)/\xi$, $\xi \in \{-10, -9.9, \ldots, 10\}$ under Gaussian error with different variances. We assumed σ_δ^2 to be known. Table 1 shows how growing measurement error variance influences the mean squared errors, averaged over 200 simulations of naive analysis, SIMEX and regression calibration. Figure 2 displays the mean prediction functions (i.e. the averaged prediction functions over 200 simulations) of these methods for $\sigma_\delta^2 = 4$. Compared to the true function and the prediction based on error free covariates, there is notable bias in all methods. However the regression calibration method outperforms the naive RVM by far.
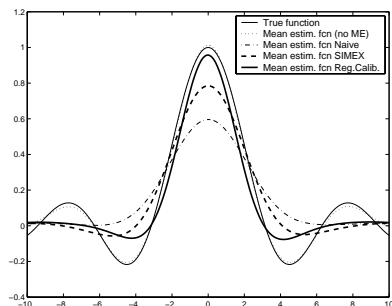
FIGURE 2. Comparison of mean prediction based on naive analysis and analysis using the true covariates without measurement error, SIMEX and regression calibration.

4   Conclusions

We see from our simulation results, how covariate measurement error invalidates the RVM regression and thus taking the error into account seems indispensable. The SIMEX and regression calibration methods seem to be able to recover the latent dependency of the target on the covariate, even under covariate measurement error.

References

Carroll, R.J., Ruppert, D., and Stefanski, L.A. (1995): *Measurement Error in Nonlinear Models.* London: Chapman & Hall/CRC.

Carroll, R.J., Maca, J.D. and Ruppert, D. (1999): Nonparametric regression in the presence of measurement error. *Biometrika*, **86**, 541-554.

MacKay, D.J.C. (1994). Bayesian non-linear modelling for the prediction competition. In: *ASHRAE Transactions, V.100, Pt.2, ASHRAE.* 1053-1062, Atlanta Georgia.

Tipping, M.E. (2000): The Relevance Vector Machine. In: *Advances in Neural Information Processing Systems.* 652-658, MIT Press.

Tipping, M.E. (2001): Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, **1**, 211-244.

Class prediction and gene selection for DNA microarrays using sliced inverse regression

Luca Scrucca[1]

[1] Dipartimento di Scienze Statistiche, Università degli Studi di Perugia, 06100 Perugia, Italy (`luca@stat.unipg.it`)

Abstract: The monitoring of the expression profiles of thousands of genes seems particularly promising for biological classification. DNA microarrays data have been recently used for the development of classification rules, particularly for cancer diagnosis. However, microarrays data present major challenges due to the complex, multiclass nature and the overwhelming number of variables characterizing gene expression profiles. We propose an approach based on sliced inverse regression which allows the simultaneous development of classification rules and the selection of those genes that are most important in terms of classification accuracy.

Keywords: Dimension reduction; SIR; Classification; Microarrays data.

1 Introduction

Gene expression data from DNA microarrays may be employed to define classification rules to predict the diagnostic category of a sample on the basis of its gene expression profile. Classification of microarray data is particularly problematic due to: (1) the large number of features (genes) from which to predict classes compared to the relatively small number of observations (samples); (2) the classification rule should be based only on those genes which contribute most to classification accuracy.

Suppose we have an expression array $\mathbf{X}$ of dimension $(n \times p)$ for n samples and p genes. The biologists view would consider $\mathbf{X}^{\top}$, in which each column represents the gene expression profile for a particular sample. We assume that gene expression measures are log transformed ratios to a baseline or a reference condition and they have already been normalized. A categorical response variable Y with K levels representing biological outcomes, such as tumors type, is also recorded along with gene expression levels. Several statistical methods have been used for classification based on gene expression profiles: discriminant analysis, logistic regression, nearest neighbor classifiers, classification trees and support vector machines (for a comparison of the above methods see Dudoit et al. (2002)).

In this paper we propose an approach based on sliced inverse regression (SIR) for class prediction and gene selection from DNA microarrays data.

We then apply the proposed methodology to a public available dataset on small round blue cell tumors (SRBCT) of childhood.

2   Applying SIR to gene expression data

Sliced inverse regression is a dimension reduction method introduced by Li (1991) which seek to find a few directions in the p-dimensional predictors space such that the regression of $Y|X$ can be fully studied on such dimension reduction subspace without loosing any relevant information contained in the data.

SIR assumes that the relationship between a response variable and a set of predictors can be expressed through the model $Y = f(\beta_1^\top X, \ldots, \beta_d^\top X, \epsilon)$, where ϵ is a random error term and $f()$ is an unknown function. The directions $(\beta_1, \ldots, \beta_d)$ span the dimension reduction subspace (drs) $\mathcal{S}(\boldsymbol{\beta})$ and must be estimated from the data. The dimension of the drs is d, and provided that the assumed model holds, we can write $Y \perp\!\!\!\perp X | \boldsymbol{\beta}^\top X$, where $\boldsymbol{\beta}$ is the $p \times d$ matrix with columns β_j. Thus, the dependence of Y on X may be fully studied through $\boldsymbol{\beta}^\top X$, the coordinates of the projection of X onto the d-dimensional subspace spanned by the columns of $\boldsymbol{\beta}$. Li (1991, Theorem 3.1) showed that, under certain conditions concerning the distribution of X, the population version of SIR is based on the following spectral decomposition:

$$\boldsymbol{\Sigma}_X^{-1} \boldsymbol{\Sigma}_{X|Y} = \mathbf{V} \Lambda \mathbf{V}^\top \tag{1}$$

where $\boldsymbol{\Sigma}_X$ denotes the covariance of X and $\boldsymbol{\Sigma}_{X|Y} = \mathrm{Var}(\mathrm{E}(X|\tilde{Y}))$, for $\tilde{Y}$ which is a sliced version of Y with fixed number of slices. Thus, the spanning matrix of the drs is given by $\boldsymbol{\beta} = \boldsymbol{\Sigma}_X^{-1/2} \mathbf{V}$. The sample version of SIR is simply obtained by replacing the above matrices with sample estimates.

Applying SIR to gene expression data appears in principle straightforward. There is no need to slice the response variable since Y is categorical with a level for each biological class. But, since $p \gg n$, $\boldsymbol{\Sigma}_X$ has rank at most n, and is hence singular and cannot be inverted (on this point see also Chiaromonte and Martinelli, 2002). However, this very large number of genes can be drastically reduced because many of them exhibit near constant expression levels across samples. A similar problem is also encountered in discriminant analysis, where it is customary to use a preliminary screening of the genes based on the ratio of between-groups to within-groups sum of squares. This statistic is clearly related to the decomposition used in computing discriminant variates, but for SIR a more natural statistic, albeit equivalent in terms of ordering, would be the ratio of between-groups to total sum of squares, i.e.

$$\frac{BSS_j}{TSS_j} = \frac{\hat{\boldsymbol{\Sigma}}_{X|Y[j,j]}}{\hat{\boldsymbol{\Sigma}}_{X[j,j]}} \qquad j = 1, \ldots, p \tag{2}$$

The $(n-1)$ genes with the largest BSS/TSS values are then selected and used to fit the SIR model. The latter provides estimates of SIR directions $\hat{\beta}_j$, $j = 1, \ldots, (K-1)$, along with the associated eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{K-1}$.

3   Class prediction and gene selection based on SIR

Expression profiles for the active genes can be projected onto the estimated drs yielding the SIR variates $\hat{\beta}_j^\top \mathbf{x}$ $(j = 1, \ldots, K-1)$. A $(K-1)$-dimensional plot using Y as marking variable may then be used to visually allocate each sample point to the closest class. A more formal procedure consists in classifying each sample to the nearest centroid in the SIR subspace. Suppose we have a test sample with expression levels $\mathbf{x}^*$, then the discriminant score for class $Y = k$ is defined as

$$\delta_k(\mathbf{x}^*) = (\hat{\boldsymbol{\beta}}^\top \mathbf{x}^* - \hat{\boldsymbol{\beta}}^\top \overline{\mathbf{x}})^\top \mathbf{W}^{-1}(\hat{\boldsymbol{\beta}}^\top \mathbf{x}^* - \hat{\boldsymbol{\beta}}^\top \overline{\mathbf{x}}) - 2\log(\pi_k) \qquad (3)$$

where the first term is the Mahalanobis distance of the test sample $\mathbf{x}^*$ with respect to the centroid on the SIR subspace, using $\mathbf{W}$ as the pooled-within class covariance matrix (which is diagonal since SIR variates are orthogonal), whereas the second term is a correction, in analogy to Gaussian LDA, based on the class prior probability, with $\sum_{i=1}^{K} \pi_i = 1$. These are usually estimated by the sample class proportions in the training data. The classification rule is then

$$\mathcal{C}(\mathbf{x}^*) = \underset{k}{\arg\min}\, \delta_k(\mathbf{x}^*) \qquad (4)$$

Discriminant scores can also be used to construct estimates of the class probabilities, i.e. $\hat{p}_k(\mathbf{x}^*) = \exp\{-\frac{1}{2}\delta_k(\mathbf{x}^*)\}/\sum_{j=1}^{K} \exp\{-\frac{1}{2}\delta_j(\mathbf{x}^*)\}$

The SIR model estimated using $(n-1)$ active genes usually provides a perfect fit to the training data, hence 0 train error rate, but it tends to be a poorer classifier for future observations. Gene selection aims at identifying a subset of genes which is able to linearly explain the patterns variation in the SIR subspace. For a two-class problem the, say, g relevant genes can be selected as those who maximizes the squared correlation coefficient:

$$R_g^2 = R^2(\mathbf{X}\hat{\boldsymbol{\beta}}, (X_{[1]}, \ldots, X_{[g]})) \qquad (5)$$

When $K > 2$ the above statistic can be generalized using the proportion $\hat{\lambda}_j / \sum_j \hat{\lambda}_j$ to reflect the importance of each estimated SIR variate. An iterative scheme is adopted: at each step only those genes which contribute the most to the overall patterns are retained and used to re-fit the SIR model. Using large values of R_g^2, say 0.999, one or few genes are removed at each step. The process is repeated until the final subset contains $K-1$ active genes. The classification accuracy of each gene subset may be assessed on the basis of its misclassification error on a test set, if available, or on a cross-validated set. This criterion may guide in choosing the "best" subset or a set of candidates subsets.

4   Classification of small round blue cell tumors (SRBCT) of childhood

We applied the above methodology to the SRBCT data provided by Khan et al. (2001). Expression measurements were obtained from glass-slide cDNA microarrays and tumors classified as Burkitt lymphoma (BL), Ewing sarcoma (EWS), neuroblastoma (NB), and rhabdomyosarcoma (RMS). 63 observations were used as training samples and 25 as test samples, although five of the latter were not SRBCTs. Khan et al. (2001) achieved a test error of 0% using a neural network approach and selected 96 genes for classification. Hastie et al. (2002) using shrunken centroids selected 43 genes, still retaining a 0% error on the test set.
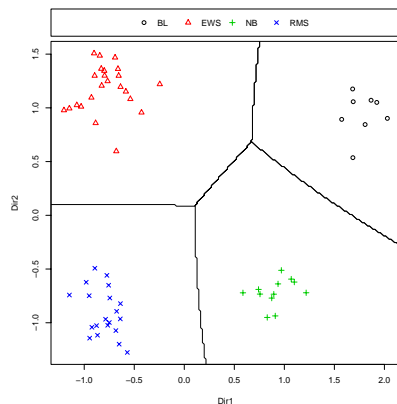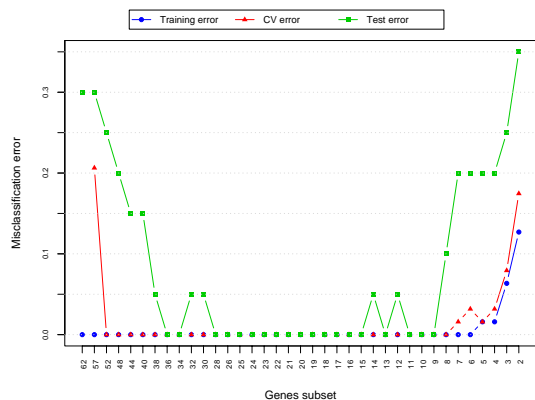
FIGURE 1. Misclassification errors for genes subsets applied to the SRBCT data.

FIGURE 2. Scatterplot of the first two SIR variates estimated using the subset of 15 genes for the SRBCT data.

Figure 1 shows the misclassification error rate for subsets of genes of decreasing size. As expected the training error appears to be an optimistic estimate of the misclassification error when compared to the test set and the CV set. From this plot we may select the subset with, say, $g = 15$ genes as the "best" subset because it has a 0 error rate on both the cross–validated and the test set. Figure 2 shows the sample points plotted on the subspace spanned by the first two SIR directions estimated using the "best" 15 genes, along with decision boundaries. The different tumor classes appear clearly separated.

Figure 3 displays the estimated probabilities each sample belonging to a given tumor class. Samples in the training set show a good separation between the highest and the next highest probability, whereas in the test set a couple of samples have less evident separation. However, even in these cases we end up with a correct classification. This kind of plot turns out to

be a very useful summary of the accuracy of the classification rule for each
sample.
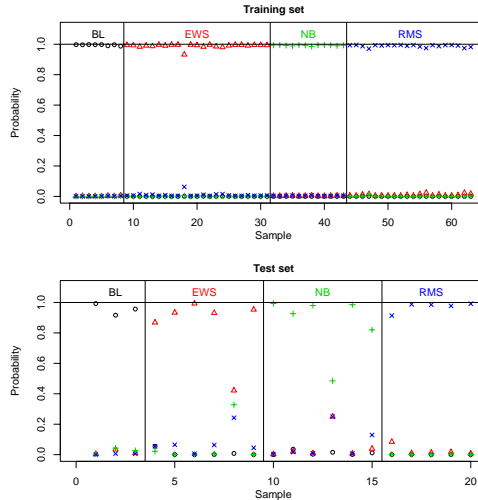
FIGURE 3. Estimated class probabilities using the "best" subset for the SRBCT
data.

References

Chiaromonte, F. and Martinelli, J. (2002). Dimension reduction strategies
for analyzing global gene expression data with a response. *Mathematical Biosciences*, **176**, 123–144.

Cook, R.D. (1998). *Regression Graphics: Ideas for Studying Regressions
through Graphics*. New York: Wiley.

Dudoit, S., Fridlyand, J. and Speed, T.P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression
data. *Journal of the American Statistical Association*, **97**, 77–87.

Khan, J. , Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F.,
Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C. and Meltzer,
P.S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, **7**, 673–679.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal
of the American Statistical Association*, **86**, 316–342.

Tibshirani, R., Hastie, T., Narashiman, B. and Chu, G. (2002) Diagnosis
of multiple cancer types by shrunken centroids of gene expression.
PNAS, **99**, 6567–6572.

Is the gene between the two markers or not?

Ib M. Skovgaard[1]

[1] Department of Natural Sciences, KVL, Thorvaldsensvej 40, DK-1871 Frederiksberg C, Denmark. E-mail: ims@kvl.dk

Abstract: A peculiar phenomenon by which the p-value for a likelihood ratio test seems to substantially exaggerate the evidence arises naturally in a genetic context. The goal of the experimenter is to map a certain recessive genetic property (mutant plant) on the chromosome. The gene has been located to the vicinity of two genetic markers, A and B say, and before further (expensive) experimentation it is useful to know whether the gene is between A and B or not. The experimental basis is the collection of all joint genotypes of A and B for all 91 mutant plants arising from a second generation cross (so-called F_2-generation) of two inbred lines. In statistical terms the example is unusual by leading from ordinary genetic probability calculations to a null-hypothesis and an alternative that are not continuously connected. Below is given some background on the modeling of the data in question.

Keywords: Gene location, markers, separate hypotheses.

1 Introduction

In the search for the location of a particular gene coding for a specific property comparison with two nearby marker genes is repeatedly used. Based on data on the three properties, the two marker types and the property in question, it is then attempted to estimate the position of the gene relative to the two markers. For the calculation of the recombination probabilities one has to distinguish whether the gene in question is between the two markers or not. This gives rise two a discontinuity in the model and makes it difficult, for example, to set up a single confidence interval for the location. Therefore it is desirable to be able to tell whether the gene is between the two markers or not. In the latter case it is usually either obvious on which side of the interval the gene is, or it is so far from the two markers that it is unimportant.

The property dealt with here is a Mendelian property, diallelic and recessive. This means that the property is governed by a single gene at which each individul has one of the two alleles, c or C, on each of the two chromosomes of the pair. The mutant allele is denoted c and the non-mutant allele is denoted C; since the property is recessive only the genotype cc leads to mutant plants, while the genotypes CC and Cc both give normal plants. The two marker genes are also diallelic but co-dominant, meaning that all

genotypes (aa, aA and AA resp. bb, bB and BB) can be distinguished. The experiment consists of sampling marker genotypes from all mutant individuals from the F_2 generation from two parent lines that are homozygotic and different on all the three genes in question. Thus, the first parental line has genotype (aa, bb, cc) and the second (AA, BB , CC). Their offspring (the F_1-generation) all have genotype (aA, bB, cC) and our plants are children of two such plants.

There were 91 mutants (genotype cc) in the F_2-generation. These were all genotyped for the two marker loci resulting in the genotype frequencies given in the following table.

Marker genotypes	bb	Bb	BB	total
aa	40	33	7	80
Aa	6	5	0	11
AA	0	0	0	0
total	46	38	7	91

The fact that the alleles a and b from the mutant parent line is much more frequent than the alleles A and B strongly suggests that the mutant gene locus is linked to the two marker loci. Further inspection clearly reveals that the a-allele is more closely linked with the mutant gene than the b-allele, suggesting that the mutant gene is closer to the A-marker than to the B-marker. Thus the order of the genes is either ACB or CAB, and our task is to estimate the distances and to distinguish the two cases.

2   Genotype probabilities

Since we observe only genotypes cc (the mutants) the observed marker genotype frequencies should be multinomial with probabilities equal to the conditional marker genotype probabilities given that the mutant locus genotype is cc. Consider first the conditional probabilities of the A-marker genotype. An AA individual implies that a cross-over has taken place for each of the two F_1-gametes, each time recombining AC and ac to Ac. Similarly a mutant of genotype Aa implies a single recombination while aa implies none. Let r denote the recombination probability between the A-marker locus and the mutant locus. Then the probabilities of the A-marker genotypes among the mutants are in Hardy-Weinberg proportions,

$$\mathrm{P}(aa|cc) = (1-r)^2, \qquad \mathrm{P}(Aa|cc) = 2r(1-r), \qquad \mathrm{P}(AA|cc) = r^2.$$

Let $X(aa)$, $X(Aa)$ and $X(AA)$ denote the observed numbers mutants with the respective genotypes. The estimate of the recombination probability, r, is then the observed recombination proportion

$$\hat{r} = \frac{2X(AA) + X(Aa)}{2n}$$

where n is the total number of mutants. I our example we get the estimate $\hat{r} = 11/(2 \cdot 91) = 0.06$. Similarly we get the estimate $\hat{s} = 0.286$ for the recombination fraction between the B-marker and the mutant locus.

When the mutant gene is between the two markers recombinations on either side of it are assumed independent and hence the conditional genotypes at the two markers are independent given the mutant (cc). Thus the nine marker genotype probabilities are obtained by multiplication of the two sets of Hardy-Weinberg proportions.

In the other case, when the gene order is CAB we no longer have this conditional independence. Then for the double heterozygotes (aA, bB) the recombination pattern cannot be completely inferred because they may arise either from the two gametes cAB and cab, or from cAb and caB. Let t denote the recombination probability between the two markers, then for this case the nine conditional probabilities are given in the following table.

Genotypes	bb	Bb	BB
aa	$(1-r)^2(1-t)^2$	$2(1-r)^2 t(1-t)$	$(1-r)^2 t^2$
Aa	$2r(1-r)t(1-t)$	$2r(1-r)(t^2+(1-t)^2)$	$2r(1-r)t(1-t)$
AA	$r^2 t^2$	$2r^2 t(1-t)$	$r^2(1-t)^2$

Whether the mutant gene is inside or outside the marker interval thus makes no difference for the conditional distributions of the two marker genotypes separately, but it does have an impact on their joint distribution, still conditioned on mutants. If we decompose our information in the mutants distribution of the A-marker genotype and the conditional distribution of the B-marker genotype given the A-marker genotype we see that the distinction between the two situations is solely in the latter component. Thus, consider the conditional distribution of B-marker genotypes given the A-marker genotype for mutants when the gene oder is CAB. This is given in the Table 3.

Genotypes	bb	Bb	BB	sum
aa	$(1-t)^2$	$2t(1-t)$	t^2	1
Aa	$t(1-t)$	$t^2+(1-t)^2$	$t(1-t)$	1
AA	t^2	$2t(1-t)$	$(1-t)^2$	1

which should be contrasted with the conditional probabilities $(1-s)^2, 2s(1-s), s^2$ for the other case. Thus, when the A-marker genotype is aa the two situations cannot be distinguished because the two conditional B-marker genotype distributions are the same except that t plays the role of s. For the genotypes Aa and AA the two conditional distributions are completely different, however. Thus it is from these two rows, in comparison with the first, that we find the information that distinguishes whether the mutant gene is inside or outside the interval.

In our data example we see that since there are no mutants with genotype AA, the 11 heterozygotes (Aa) distributed as (6,5,0) on (bb, Bb, BB) are

crucial. As it turns out their distribution is in perfect accordance with the mutant gene being between the two markers, but does not fit quite well with the other situation by which, among other aspects, bb and BB should be equally likely. The problem is, however, whether the information is sufficient to exclude this situation with reasonable degree of certainty.

3   Testing one situation against the other

For the purpose of excluding one of the situations we would like to set up a test for this situation, with the other as alternative, and vice versa. Using the conditional distribution of the B-marker given the A-marker for this inference we obtain two mathematically disconnected models, each parametrized by a single parameter, either t or s from above, each varying between zero and a half. Actually the two models have a single point in common, corresponding to $t = 0.5$ and $s = 0.5$. This is the situation when there is no linkage between the mutant gene and any of the markers, thus contradicting that the mutant gene is between the two markers. Mathematically this contradiction is removed if we limit s upwards by the recombination fraction between the two markers. This is unimportant for our case since our interest is rather at the other end of the distribution with s or t near zero.

There are (at least) three methods at hand for the present case. One is to make a goodness-of-fit chi-squared test based on expected numbers under the two hypotheses. This gives the p-value 0.79 for the hypothesis that the gene is between the markers, and $p = 0.036$ for the other hypothesis. But this is a weak test since it tests against any alternative without taking advantage of our genetic knowledge.

The second possibility is to use a likelihood ratio test for each of the to hypotheses against the other. Using the asymptotic distribution (Cox, 1962) we get the two p-values 0.66 and 0.0004, this time seemingly giving overwhelming evidence that the gene is between the two markers.

However the likelihood ratio itself is only 0.028 giving posterior odds around 35 in favor of the gene being between the markers based on equal prior probabilities. Although the conclusion is in the same direction as with the other tests, the likelihood ratio and the posterior odds suggest that the p-value exaggerates the evidence by a factor around 10 in this example.

References

Cox, D.R. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society, Series B*, **24**, 406–424.

Bayesian Covariance and Variable Selection for Explaining Consumer Behaviour

Regina Tüchler[1]

[1] Dept. of Statistics and Mathematics, University of Economics and Business Administration, Augasse 2-6, A-1090 Vienna, Austria, e-mail: regina.tuechler@wu-wien.ac.at

Abstract: We estimate the random coefficient model by means of Markov Chain Monte Carlo methods (MCMC) and simultaneously carry out variable selection and covariance selection during our modeling procedure. Following the statistical principle of parsimony this method yields a model, which includes only the significant variables and covariance elements and therefore allows a more efficient estimation. It offers a reasonable basis for making decisions in real applications. We will demonstrate this for marketing data which come from conjoint analysis. In this application the heterogeneous behaviour of consumers has to be explained from high-dimensional data.

Keywords: Covariance Selection, Variable Selection, Random Coefficient Model, MCMC Methods, Heterogeneity Model

1 The Model

The procedure of this paper is based on the following random coefficient model:

$$y_i = Z_i \Theta z_i + Z_i \beta^G + Z_i C \tilde{z}_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2 I), \qquad (1)$$
$$\tilde{z}_i \sim N(0, I), \qquad (2)$$

where I denotes the identity matrix. We have T_i observations y_i for each subject $i = 1, \ldots, N$. Z_i are the design matrices of dimension $T_i \times d$. We include r covariates z_i into the model and the $d \times r$-dimensional matrix Θ is the corresponding parameter matrix. C is a lower triangular squared matrix and $\tilde{z}_i$ are standard normally distributed. (1), (2) is equivalent to the following traditional representation of a random coefficient model:

$$y_i = Z_i \Theta z_i + Z_i \beta_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2 I), \qquad (3)$$
$$\beta_i = \beta^G + u_i, \quad u_i \sim N(0, Q = CC'). \qquad (4)$$

The lower triangular matrix C is the Cholesky factor of the Cholesky decomposition of the covariance matrix of the random effects Q.

Usually we do not have additional prior information about the selection of variables and the form of the covariance matrix. Therefore on the one hand it is desirable to start with a very general model involving all covariates and all effects as random effects. On the other hand the estimation of a large parameter vector with possibly unnecessarily many elements reduces the efficiency and the speed of convergence of the MCMC chains. To deal with both these aspects we formulate our model in a general way and let the data choose the special structure during the modeling procedure. Therefore we add indicators δ and γ to our model parameters. These indicators define which elements of Θ and C are excluded from the estimation:

$$
\begin{array}{ll}
\Theta_{jk} = 0, & \text{iff } \delta_{jk} = 0, \\
\Theta_{jk} \neq 0, & \text{iff } \delta_{jk} = 1,
\end{array}
\quad \text{and} \quad
\begin{array}{ll}
C_{lm} = 0, & \text{iff } \gamma_{lm} = 0, \\
C_{lm} \neq 0, & \text{iff } \gamma_{lm} = 1, \\
\text{for } l \geq m.
\end{array}
\tag{5}
$$

Only those elements of Θ and C which are unequal to zero are included into the estimation procedure and are denoted Θ^δ and C^γ, respectively. Bayesian estimation via MCMC methods amounts to the estimation of the unknown model parameters $\Theta^\delta, \beta^G, C^\gamma, \sigma_\varepsilon^2$ together with the indicators γ and δ and augmented by the individual effects $\tilde{z}_i$.

2   Bayesian Estimation Using MCMC Methods

2.1   MCMC Sampling Steps for the Parsimonious Estimation of Random Coefficient Models

The following MCMC steps are involved:

(I) Generate from $\delta_{jk}|\delta_{\setminus jk}, \gamma, \beta^G, \tilde{z}, \sigma_\varepsilon^2, y$.

(II) Generate from $\gamma_{lm}|\gamma_{\setminus lm}, \delta, \beta^G, \tilde{z}, \sigma_\varepsilon^2, y$.

(III) Generate from $\Theta^\delta, C^\gamma|\gamma, \delta, \beta^G, \tilde{z}, \sigma_\varepsilon^2, y$.

(IV) Generate from $\beta^G|\Theta^\delta, C^\gamma, \sigma_\varepsilon^2, y$.

(V) Generate from $\tilde{z}|\beta^G, \Theta^\delta, C^\gamma, \sigma_\varepsilon^2, y$.

(VI) Generate from $\sigma_\varepsilon^2|\beta^G, \Theta^\delta, C^\gamma, \tilde{z}, y$.

We denote the data y and the individual effects $\tilde{z}$ for all subjects i. $\delta_{\setminus jk}$ is the notation used for the sequence δ excluding δ_{jk} and similarly for $\gamma_{\setminus lm}$. Steps (IV), (V) and (VI) are standard MCMC steps described for example in Frühwirth-Schnatter *et al.* (2004). In step (I) and (II) the indicators are generated applying the efficient sampling scheme of Smith and Kohn (2002). In step (III) we generate Θ^δ and C^γ jointly from a multivariate normal distribution.

2.2   Comparison to Existing Estimation Procedures

Usually model (1), (2) is estimated for the unrestricted parameter matrices
Θ and C. One common way is to center the random effects according to
the transformation (4) and to assume a prior inverted Wishart distribution
for the covariance matrix Q. For such an algorithm the choice of the prior
parameters has a big influence on the estimates and on the speed of conver-
gence (Natarajan, Kass 2000). Furthermore it includes the strong prior as-
sumption of a full covariance matrix. But a prior determination of non-zero
variances in Q is only reasonable, if we are sure that we really have *random*
effects. If the decision about the effects being fixed our random is uncertain
such an algorithm may yield an overfitted model, including unnecessarily
many effects as random. Additionally we may also exclude non-significant
covariances even if the corresponding variances are unequal to zero for our
algorithm. Therefore the covariance matrix may be estimated in a more
flexible way than for other methods (e.g. Chen, Dunson (2003)). Similar
arguments are true for the estimation of the parameters Θ. In real applica-
tions we typically have a huge number of variables, but many of them are
likely to be zero. Including all of them is unsatisfactory. An advantage of
our procedure is that it does not involve a prior decision about the form
of the covariance matrix and the selection of the variables. These decisions
are made based on the data during the modeling procedure.

3   Estimation of Heterogeneity in the Mineral Water Market

The data of our application come from the Austrian mineral water market
and have already been estimated by means of a traditional Gibbs sampler
at the IWSM (Frühwirth, Otter 1999 and Tüchler *et al.* 2002). The design
matrices Z_i consist of the following 15 columns: 7 main effects (constant,
4 brands, price and quadratic price), 4 brand by price and 4 brand by
quadratic price interaction effects. 213 consumers stated their likelihood to
buy 15 different mineral water products on a 20 point rating scale. This
yields a design matrix of dimension 15×15. Additionally we include 7 con-
sumer characteristics z_i into the analysis. We have 120 distinct elements
in the covariance matrix Q and also 120 elements in the parameter ma-
trix Θ. The dimensions of the parameters in this application are big and
advantages of parsimonious estimation of Q and Θ are to be expected.
¿From the marketing point of view the following questions are of interest.
Do the consumers behave homogeneously with respect to some of the effects
of Z_i? Are there dependencies between those effects for which we found a
heterogeneous behaviour? Do consumer specific attributes really help to
understand the consumer market in the mineral water category?
To answer all these questions we look at posterior estimates of the indica-
tors δ and γ. These may be interpreted as probability of an element of Θ

TABLE 1. Posterior probability for the elements of the covariance matrix Q to be significantly different from zero.

1	1	1	1	1	1	1	1	1	.0	.6	0	0	.0	0
-	1	1	1	1	1	1	1	1	.0	.6	1	1	.0	.0
-	-	1	1	1	1	1	1	1	.0	.6	1	1	.0	.0
-	-	-	1	1	1	1	1	1	1	1	.0	.0	.0	.0
-	-	-	-	1	1	1	1	1	.0	1	.0	.0	.0	.0
-	-	-	-	-	1	1	1	1	.9	.6	1	1	.0	.0
-	-	-	-	-	-	1	1	1	.0	1	.2	.2	.0	.0
-	-	-	-	-	-	-	1	1	1	.6	1	.0	.0	1
-	-	-	-	-	-	-	-	1	1	.6	.0	.1	.0	.0
-	-	-	-	-	-	-	-	-	1	.0	.0	.1	.0	.0
-	-	-	-	-	-	-	-	-	-	1	.0	.0	.0	.0
-	-	-	-	-	-	-	-	-	-	-	1	1	.0	1
-	-	-	-	-	-	-	-	-	-	-	-	1	.0	.0
-	-	-	-	-	-	-	-	-	-	-	-	-	.0	0
-	-	-	-	-	-	-	-	-	-	-	-	-	-	1

and C, respectively to be non-zero. The posterior probability for the elements of the covariance matrix Q to be non-zero are given in Table 1. Only the interaction effect of one brand by the quadratic price has a low posterior probability of 0.04 for being a random effect. Our algorithm estimates this effect as a fixed effect as we can see from the zeros in the fourteenth column of Table 1. For all other effects the indicators clearly advocate for treating them as random. In Table 1 the diagonal elements take values of one for these effects. The correlation between the different random effects is clearly present for the 7 main effects (the value of the indicators is one), whereas this probability is rather close to zero for many of the interaction effects. Note that for our selection algorithm it is possible to include the variances of these interaction effects into the model whereas the non-significant covariances are ignored. Here the new procedure offers interesting results in comparison to earlier model selection for these data. In Tüchler *et al.* (2002) we chose a model with fixed brand by quadratic price interactions for all four brands. So another three effects were fixed. Since we estimated the covariance matrix from an inverted Wishart distribution, the decision between fixed or random brand by quadratic price interactions involved the decision about 54 additional elements and the model with fewer parameters was preferred then. For our new procedure we decide for each element separately and the flexibility of this methods allows to select 13 elements of the covariance matrix for the brand by quadratic price interactions (see the last four columns in Table 1).

Looking at the posterior probabilities of the parameters Θ to be non-zero

we find that the consumer specific variables do not deliver much additional insight in the behaviour of consumers in the mineral water market. Only two effects concerning the education and the income have an indicator with posterior probability of 1 for being unequal to zero. For all others we obtain a probability between 0 and 0.15. This is in line with marketing theory that says that consumer specific attributes are unimportant for the explanation of heterogeneity in such a market of convenience goods.

References

Chen, Z. and Dunson, D. B. (2003). Random Effects Selection in Linear Mixed Models. *Biometrics*, **59**, 762-769.

Frühwirth-Schnatter, S. and Otter, Th.] (1999). Conjoint-Analysis Using Mixed Effect Models. In: *Statistical Modelling. Proceedings of the Fourteenth International Workshop on Statistical Modelling*, ed. Friedl, H., Berghold, A. and Kauermann, G., Graz, pp. 181-191.

Frühwirth-Schnatter, S., Tüchler, R. and Otter, Th. (2004). Bayesian Analysis of the Heterogeneity Model. *Journal of Business and Economic Statistics*, **22**, 2-15.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian Variable Selection. *Statistica Sinica*, **7**, 339-373.

Natarajan, R. and Kass R. E. (2000). Reference Bayesian Methods for Generalized Linear Models. *Journal of the American Statistical Association*, **95**, 227-237.

Smith, M. and Kohn R. (2002). Parsimonious Covariance Matrix Estimation for Longitudinal Data. *Journal of the American Statistical Association*, **97**, 1141-1153.

Tüchler R., Frühwirth-Schnatter S. and Otter, Th. (2002). The Heterogeneity Model and its Special Cases - an Illustrative Comparison. In: *Proceedings of the 17th International Workshop on Statistical Modelling*, ed. Stasinopoulos, M., and Touloumi, G., Chania, Greece, pp. 637-644.

Hierarchical Bayesian Modelling of Spatial Interactions of Gene Expression on the Tuberculosis Genome

Ernst Wit[1], Nial Friel[1]

[1] Department of Statistics, University of Glasgow

Abstract: *M. Tuberculosis* is a bacterium with a ring-shaped genome. Microarray experiments make it possible to monitor the gene expressions of each of the 3924 genes over time. Given that the genes are so tightly packed on the genome, it is expected that neighbouring genes influence each other. We define a Hidden Markov Model (HMM) to relate the observed expression levels to hidden states "Up", "Down" and "Same" for a time-series gene expression dataset with four time points. A Potts model is identified to describe the interactions between neighbouring states. A typical problem in these types of model is the estimation of the parameters of the hidden states because of the intractability of the normalizing constant. Recent work by Pettitt et al. (2003) provides a clue to avoid using a pseudolikehood approximation.

Keywords: microarray; hidden Markov model; gene interaction; normalizing constant.

1 Introduction

Microarray technology has made the simultaneous measurement of gene transcription a routine activity. Whereas gene transcription is only one stage in the complex genomic process of living organisms, it gives a fascinating insight in one aspect of this activity across the whole genome.
Gene regulation is a complex biological process which involves gene-gene and gene-protein interactions. Some of the interactions may be on a local scale. A particular strand of *Mycobacterium Tuberculosis* has a genome with 4,411,529 base pairs, on which 3,924 genes are rather tightly packed. If during the process of transcription, the RNA polymerase enzyme, by chance, skips the inhibitor and the neighbouring genes are in the same direction, then it might be the case that neighbouring genes tend to be co-expressed. Figure 1 shows this co-expression hypothesis. A similar hypothesis was put forward in Oliver et al. (2002). In this paper we describe a model to analyze the hypothesis for positive local interactions between genes.

2 Time-course gene-expression experiment

Prof. Phil Butcher and his Bacterial Microarray Group (Bμgs) at St. George's Hospital in London are interested in studying the effects of stressed growth
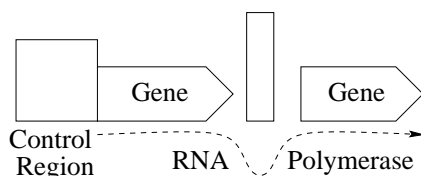
FIGURE 1. As genes are closely packed in a M. Tuberculosis genome, it is not unlikely that the RNA polymerase enzyme skips the inhibitor and also expresses the subsequent gene.

on the expression levels of all 3924 genes in *M. tuberculosis*. In one experiment, five cultures of *M. tuberculosis* were grown with only a limited growth medium. The cultures in the first two flasks were grown until day 6 and then harvested. The other three cultures were grown and harvested at day 14, 20 and 30, respectively. From each harvest four batches of RNA were extracted and hybridized to four microarrays with a genomic DNA reference sample.

Although it is possible to model the quantitative expression data in a continuous fashion, there are two reasons why it is more satisfactory to model the data in a discrete way:

1. There is biological evidence that the biological relevance of differential expression is unrelated to its associated fold-change (Johnson et al. 2003). A small fold-change can have the same effect as a large fold-change.

2. As a consequence of the noisy nature of gene expression data with many outliers, modelling discrete interactions are more robust.

For this reason, we define the hidden states—"down" (-1), "same" (0) and "up" ($+1$)—and define the spatial interactions between these states. Conditionally on the hidden states, we define the likelihood of the data.

3   Interaction Model

Each microarray contains 4,624 spots, among which 3,924 are *M. Tuberculosis* genes. For each of the genes, the position on the *M. Tuberculosis* genome is known. Like many bacteria, the genome of *M. Tuberculosis* is circular. This means that the last gene, Rv3924, is right next to the first, Rv0001. The expression of the genes is observed over four time-points, i.e., across three transitions. The underlying structure of the data, therefore, can be described as a $3,924 \times 3$ cylinder s, as shown in Figure 2.
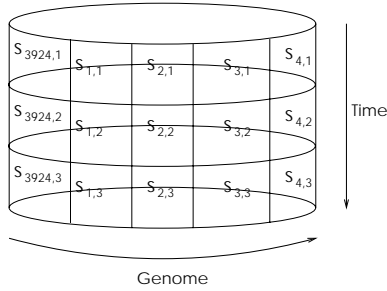
FIGURE 2. The hidden parameter s defines a Markov Random Field on a lattice that wraps around.

3.1   Hidden Potts model for gene interactions

The hidden states $s_{ij} \in \{-1, 0, +1\}$ form a discrete lattice, on which we define spatial and temporal interactions. Potts models have typically been used for these kinds of purposes. Our model is in spirit close to such Potts models, except that it explicitly takes into account the ordered nature of the states,

$$p(s_{ij}|s_{-ij}\ \theta_m) \quad \propto \exp \quad (\theta_t \sum_{m \sim i} \frac{2 - |s_{mj} - s_{ij}|}{2} + \theta_g \sum_{n \sim j} \frac{2 - |s_{in} - s_{ij}|}{2}$$
$$+ \theta_{-1} 1_{\{s_{ij}=-1\}} + \theta_0 1_{\{s_{ij}=0\}}), \tag{1}$$

where $m \sim j$ and $n \sim i$ refer to neighbouring cells in the vertical and horizontal direction, respectively, keeping in mind the cylindrical nature of the lattice. The parameters θ_t and θ_g describe the interactions in the time and spatial, i.e. genome, components. Positive values of these parameters make it more likely that the same state persists across time and across the genome, whereas negative values of θ_t and θ_g increase the likelihood of opposite states.

3.2   Likelihood of the data

The idea is to define the likelihood of the data conditional on the hidden states. Rather than considering the full 3924×16 data matrix, we only consider a summary thereof, which, although not sufficient, is, in some approximate sense, "close" to such. For evaluating mean changes across two populations, the t-statistic is most powerful, if the underlying data are normally distributed. For this reason, we define for each gene g three t-statistics across time,

$$d_{gi} \quad = \quad \frac{\bar{x}_{g,i+1,.} - \bar{x}_{gi.}}{s_{p(i,i+1)}}, \quad i = 1, 2, 3, \tag{2}$$

where the expression values x_{gij} are considered on the log-scale, which can be assumed approximately normal (Wit and McClure 2004). Conditional on the hidden states, the vector $(d_{g1}, d_{g2}, d_{g3})^t$ has a multivariate t-distribution with a known covariance structure. The non-centrality parameters are assumed to be fixed, $\mu_{-1} < 0$, $\mu_0 \equiv 0$, $\mu_{+1} > 0$, and depend on whether a hidden state is -1, 0 or $+1$ respectively.

4   Model estimation via MCMC

By putting priors on all the model parameters, the model is a typical Bayesian hierarchical model and most of the parameters can be updated rather standardly via Gibbs or Metropolis-Hasting procedures. The parameters of the hidden interaction model θ are an exception, because the likelihood $p(\theta|s, d, \mu)$ is only defined up to a normalizing constant that itself depends on θ. Usually, this is remedied by using the pseudo-likelihood, but recent work by Pettitt et al. (2003) make it possible to calculate the normalizing constant exactly.

Theorem. *Let $s = (s_1, s_2, \ldots, s_n)$ a cylindrical lattice with n columns, and let $q(s|\theta) = \prod_{i=1}^n h_\theta(s_i, s_{i+1})$ the unnormalized density on s, where h_θ is a homogeneous transfer function, then the normalizing constant of $q(.|\theta)$ is given by*

$$Trace\left(Q^n\right),\tag{3}$$

where Q is a $N \times N$ matrix, defined via $Q_{kl} = h_\theta(s_1 = a_k, s_2 = a_k)$, where $A = \{a_1, a_2, \ldots, a_N\}$ the set of all values a column of s can assume.

In our case, the transfer function h_θ is given by

$$h_\theta(s_i, s_{i+1}) = \exp\left(\theta_t \sum_{j=1}^{2} \frac{2 - |s_{ij} - s_{i,j+1}|}{2} + \sum_{j=1}^{3} [\frac{2 - |s_{ij} - s_{i+1,j}|}{2} \, \theta_g \right.$$
$$\left. +1_{\{s_{ij}=-1\}} \, \theta_{-1} + 1_{\{s_{ij}=0\}} \, \theta_0]\right).\tag{4}$$

In each MCMC sweep this quantity has to be calculated. The 27×27 matrix Q has only positive entries, is therefore irreducible and by the Perron-Frobenius theorem can be partitioned $Q = H^{-1}DH$, whereby D is a diagonal matrix. The normalizing constant is therefore easily calculated as $\text{Trace}[D^{3924}] = \sum_{i=1}^{27} D_{ii}^{3924}$. The computational effort is thus exactly the same as for pseudo-likelihood.

5   Results

The sampler was initialized by reasonable values for each of the parameters. Figure 3 seems to suggest that the sampler burned in and converged
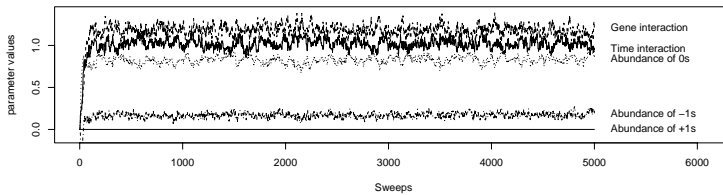
FIGURE 3. MCMC run of the parameter estimates for θ.

relatively quickly. This was confirmed by choosing different starting points for each of the parameters and redoing the sampler (results not shown). The posterior mean of θ_g is 1.12, which suggests a positive relationship between neighbouring genes. This confirms the hypothesis that control mechanisms of a simple organism such as a *M. Tuberculosis* bacterium have a local component, which leads to co-expression of neighbouring genes. The parameter θ_t is also positive, suggesting that gene expression changes tend to persist in time. Moreover, the abundance parameter for state 0 is quite a bit larger than the abundance parameters for states -1 and $+1$, i.e., most of the genes don't change expression level most of the time.

References

Pettitt, A.N., Friel, N., Reeves, R. (2003) Efficient calculation of the normalisation constant of the autologistic model on the lattice. *Journal of the Royal Statistical Society, Series B*, **65**:1, pp. 235-247.

Johnson, C.D., Balagurunathan, Y., Lu, K.P., et al. (2003) Genome profiles and predictive biological networks in oxidant-induced atherogenesis. *Physiol. Genomics*, **13**, pp. 263-275.

Oliver B., Parisi M., Clark D. (2002) Gene expression neighborhoods. *Journal of Biology*, **1**:1, article 4.

Wit, E.C., McClure, J.D. (2004). *Statistics for Microarrays; Design, Analysis and Inference.* Chichester: John Wiley & Sons.

Poster Sessions

Multivariate Linear Model for Selection of Oilseed Rape Genotypes

Zygmunt Kaczmarek[1], Elżbieta Adamska[1] and Teresa Cegielska-Taras[2]

[1] Institute of Plant Genetics, Polish Academy of Sciences 60-479 Poznań, ul. Strzeszyńska 34, Poland, e-mail: zkac@igr.poznan.pl
[2] Plant Breeding and Acclimatization Institute, Poznań, Poland

Abstract: Utilizing the multivariate analysis of variance approach it is shown how doubled haploids lines of oilseed rape can be selected with respect to the content of favorable fatty acids. Investigation the way of the genetic improvement and selection forms characterized both a higher oleic acid and the ratio of linolenic and linoleic acid (1:2).

1 Introduction

Winter rapeseed (*Brassica napus* L.) become a major oilseed crop in Europe when quality varieties, low in erucic acid and glucosinolate content were developed and introduced into commercial production. Quality improvement in both the oil and meal portion of the seed were key factor in the success of rapeseed as a new, high quality and edible oil.
Fatty acid composition of the zero erucic acid commercial *Brassica napus* L. crop is typical for this species and similar to what observed in the past over many years. Rapeseed oil has high concentration of oleic acid (about 60%), and contains moderate levels of linoleic acid (about 20%) and linolenic acid (about 10%). This fatty acid composition of a vegetable oil is considered ideal by many nutritionists for human nutrition, and superior to that of many other plants oils. Rapeseed oil also has the lowest saturated fatty acid of any vegetable oil of about 7% of total fatty acids, whereby palmitic acid (C16:0) with about 4% and stearic acid (C18:0) with about 2% of the total fatty acids, are the major saturated fatty acids in rapeseed oil. But reduced levels of the polyunsaturated fatty acids, such as linolenic acid (C18:3), and increased levels of the monounsaturated oleic acid (C8:1) are associated with higher oxidative stability.
During last two decades tremendous progress has been made in the *in vitro* production of haploid plants. Rapeseed is species where doubled haploids

(DH) are produced with high efficiency and the system is widely applied in breeding.

For improving the breeding efficiency, the selection of oilseed rape genotype according to the desirable fashion e.g. regarded as nutritionally favorable of fatty acids composition the study was taken. The primary objective of this study was to investigate the way of the genetic improvement and selection forms characterized both a higher oleic acid and the ratio of linolenic and linoleic acid (1:2).

2   Description of the data

Two doubled haploid (DH) lines of winter oilseed rape, DH-0120 (P1) and DH-C1041 (P2) were crossed to produce a hybrid generation, F1. The F1 gametes were sampled to develop doubled haploid population using the isolated microspore culture method (Cegielska-Taras et al. 1997).

In this paper the analysis of results of experiment with 32 doubled haploids, 2 parental forms P1 and P2 and oilseed rape standard variety Kana, conducted at one place in 2000, is presented. The content of following acids was observed and analysed: palmitic acid (C16:0), stearic acid (C18:0), oleic acid (C18:1), linoleic acid (C18:2) and linolenic acid (C18:3). The data analysed here form a part of a much larger research project concerning the breeding and selection of oilseed rape genotypes. Therefore, only some results of the basic analysis will be shown. But before that, the model adopted for the analysis is to be specified.

3   Mathematical model of observations

The data coming from the experiments with rapeseed genotypes are multivariate, because they originate from measurements taken on a set of mutually interrelated characteristics. A method which takes into account the interrelation between various acids is the multivariate analysis of variance (MANOVA).

Let $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n]'$ be the matrix of n observations of p quantitative traits such that $E(\mathbf{Y}) = \mathbf{X}\mathbf{\Xi}$, where $\mathbf{X}$ is the $n \times q$ design matrix of rank $r \leq q$, and $\mathbf{\Xi} = [\xi_1, \xi_2, ..., \xi_p]$ is the $q \times p$ matrix of unknown parameters. Vectors $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n$ are p-dimensional observations, each having an independent normal distribution with the same unknown nonsingular covariance matrix $\mathbf{\Xi}$, i.e. each $\mathbf{y}_i$ $(i = 1, 2, ..., n)$ is distributed independently according to $N[E(\mathbf{y}_i), \mathbf{\Sigma}]$. Then the p-variate MANOVA model may be written in form

$$\mathbf{Y} = \mathbf{X}\mathbf{\Xi} + \mathbf{E}, \tag{1}$$

where $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n]'$ is the matrix of errors with $\mathbf{e}_i \sim N(0, \mathbf{\Sigma})$ for all i.

4   Tests of hypotheses

In the analysis of multivariate experimental data, the interest may be in testing some hypotheses of the type

$$H_o : \mathbf{C\Xi M = 0}, \tag{2}$$

where the $g \times q$ matrix $\mathbf{C}$ is of rank g and the $p \times u$ matrix $\mathbf{M}$ is of rank u. The rows of $\mathbf{C}$ represent a set of contrasts between the q rows of $\mathbf{\Xi}$ and the columns of $\mathbf{M}$ represent some combinations of the columns of $\mathbf{\Xi}$ which correspond to the observed variables. The necessary and sufficient condition for H_o to be testable is the equation $\mathbf{C(X'X)^- X'X = C}$, where $\mathbf{(X'X)^-}$ is a generalized inverse of the matrix $\mathbf{X'X}$. Thus, the hypothesis H_o may be tested with any of the following test statistics (cf. Morrison 1976): the Wilks likelihood ratio Λ, the Hotelling-Lawley trace T_o^2, the Pillai trace V or the Roy maximum characteristic root $c_{\max}$. Any of above tests involves the computation of the two matrices: the sum of squares of products matrix for error

$$\mathbf{S}_E = \mathbf{M'Y'(I}_n - \mathbf{X(X'X)^- X')YM}, \tag{3}$$

and the matrix for hypothesis

$$\mathbf{S}_H = \mathbf{M'YX(X'X)^- C'[C(X'X)^- C']^{-1}C(X'X)^- X'YM}. \tag{4}$$

To test the hypothesis H_o it will be convenient to use the Hotelling-Lawley trace statistic defined as

$$T_o^2 = (n - r)\text{trace } (\mathbf{S}_E^{-1}\mathbf{S}_H) \quad \text{(Lejeune, Caliński, 2000).} \tag{5}$$

The critical values at the significance level α, equal $T_{o,\alpha,u,g,n-r}^2$, were given by Seber (1984). However a suitable F-test approximation defined by Mc Keon (1974) is available and will be used in this paper.

If H_o is rejected, one may be interested in testing hypotheses implied by H_o, particularly

$$H_{i,0} : \mathbf{c}_i'\mathbf{\Xi M = 0'}, \quad H_{0,j} : \mathbf{C\Xi m}_j = \mathbf{0}, \text{ and } H_{i,j} : \mathbf{c}_i'\mathbf{\Xi m_j} = 0, \tag{6}$$

when matrices $\mathbf{C}$ and $\mathbf{M}$ are replaced by row $\mathbf{c}_i'$ and column $\mathbf{m}_j$ correspondently for all i and j ($i = 1, 2, ..., g; \ j = 1, 2, ..., u$).
The appropriate Hotelling-Lawley statistics for testing these hypotheses are known (Lejeune, Caliński, 2000).

5   Analysis of the data

As mentioned in Section 2 the data come from the experiment in which 35 genotypes of winter rapeseed were compared with respect to five fatty

TABLE 1. Estimates and results of testing the contrasts with cv. Kana for the selected DH lines

Contrast DH line – cv.Kana	DH line Nr.	Estimate of contrast in the fatty acid			F-value for multivariate test
		C16:0+C18:0	C18:1	C18:2 − 2×C18:3	
H5–30	2	-0.27	2.97*	-0.17	2.87
H5–43	3	-0.40	1.97	-0.03	0.97
H5–109	9	-0.17	1.13	-3.73	3.75
H5–129	11	-0.23	3.37**	0.40	3.33
H5–255	18	-0.47	1.43	0.20	0.80
$F_{0.05}$					2.74

*,** – denotes statistical significance at the level of 0.05 and 0.01 respectively

acids. The experiment was conducted in a completely randomized block design with $r = 3$ replications. The experimental data from the $n = 105$ plots are multivariate as the observations were taken on $p = 5$ variables (fatty acids). The experiment was analysed under the usual model for a block design, which in the multivariate case can be written in accordance with (1). The data were analysed with respect to two aspects: the proper selection of DH lines and estimation of transgression effects of doubled haploids, for oleic acid.

In order to select the best lines in terms of requirements described in introduction, it was suggested – using the basic results of MANOVA performed for the five analysed acids – to test the contrasts of the individual DH lines with the standard, taking into consideration three "combinations of variables" being the functions of the analysed acids. These variables were defined to meet the assumed requirements of the line evaluation.

Thus, the first variable concerns the total saturated acids (C16:0 + C18:0), the second – the content of oleic acid (C18:1), the third – the difference between linoleic acid and the doubled content of linolenic acid (C18:2 – 2 x C18:3). Cultivar Kama turned out to be suitable as a standard as it exhibited an almost exactly 2:1 ratio of the linoleic (21.20%) to linolenic acid (10.57%) contents at the oleic acid content 61.37%. In this purpose the appropriate hypotheses given in (2) and (6) were tested taking as a columns of matrix $\mathbf{M}$ three vectors: $\mathbf{m}_1 = [1\ 1\ 0\ 0\ 0]'$, $\mathbf{m}_2 = [0\ 0\ 1\ 0\ 0]'$ and $\mathbf{m}_3 = [0\ 0\ 0\ 1\ -2]'$ and as a $\mathbf{c}_i'$ the vectors of coefficients equal to 1 for ith line, -1 for cv. Kana and zero for the rest lines.

The results of testing above mentioned hypotheses allowed to reject the general hypothesis H_0 of no differences between DH lines with regard to three new variables ($F = 3.94 > F_{0.01} = 1.48$). It was shown that five doubled haploid lines had a higher content of oleic acid than cv. Kana and almost exactly 2:1 ratio of linoleic to linolenic acids. However, only for two lines H5-30 and H5-129 the difference in the oleic acid content was positive and significant (at $\alpha = 0.01$ and $\alpha = 0.01$ respectively). The results of evaluated five selected DH lines are given in Table 1.

An additional comparison of the individual DH lines with the mean of parental forms makes it possible to assess the transgression effects of these lines in terms of the oleic acid content. The results of evaluation of these effects indicate the occurence of transgression in seven DH lines.

References

Cegielska-Taras T., Szała L. (1997). Plant regeneration from microspore derived embryos of winter oilseed rape (*Brassica napus* L.). *Oilseed Crops*, XVIII, 21-30.

Lejeune M., Caliński T. (2000). Canonical Analysis Applied to Multivariate Analysis of Variance. *Journal of Multivariate Analysis* 72, 100-119.

McKeon J.J. (1974). *F* approximations to the distribution of Hotelling's T_0^2. *Biometrika* 61, 381-383.

Morrison D.F. (1976). *Multivariate Statistical Methods.* 2nd ed., McGraw-Hill, New York.

Seber G.A.F. (1984). *Multivariate Observations.* Wiley, New York.

Mixed Model for Studying the Stability of Phenotypic Gene Effects

Maria Surma[1], Zygmunt Kaczmarek[1], Tadeusz Adamski[1]

[1] Institute of Plant Genetics, Polish Academy of Sciences 60-479 Poznań, ul. Strzeszyńska 34, Poland, e-mail: zkac@igr.poznan.pl

Abstract: The statistical model for the experiments repeated with the same set of genotypes at several locations over a period of years is presented. The model has been defined for the statistical analysis of experiments with the same set of genotypes conducted in the completely standarized block design. The methodology of analysis the data from such a series of experiments was applied to the study of gene effects on the basis of doubled haploids population and F_2 and F_3 hybrid generations. Practical application of this approach was shown on an example concerning the interaction of gene effects with environments for coarse extract yield of barley.

Keywords: gene effects; stability; continuous variables; multivariate analysis; F-statistic; GE interaction

1 Introduction

Information on genetic determination of quantitative traits may be obtained by estimation of genetic parameters,connected with gene effects, on the basis of phenotypic observations. Such estimation may be made on the basis of early generation (Mather, Jinks, 1982), or on a population of doubled haploids (DH) lines derived from F_1 hybrids of two homozygous parents (Surma et al., 1997). In both cases estimators of the parameters are some functions of mean of the studied generation. Phenotypic values of traits are conditioned by both genetic and environmental factors. The problem is more complicated when genotype-environment interaction occurs; it may greatly influence the differences between the studied generations, and consequently estimates of the genetic parameters. Therefore, to obtain credible information on inheritance of metrical traits, the GE interaction should be taken into account in the genetic analysis. Especially important is information concerning stability of phenotypic gene effects. Methods of statistical analysis of a series of genetic experiments given by Caliński et al. (1997) permit to evaluate the interaction with environments for each genotype. Estimation of stability is based on GE interaction effect related to each genotype measured be the value of the relevant F-statistic. Similarly, phenotypic gene effect can be recognize as a stable when F-statistic

value for its interaction with environments is at less than critical value.

2   Description of the data

Thirty doubled haploid lines of barley, derived from F_1 hybrids of the malting cultivar Grit and the non-malting cultivar Havila were used in the study (Kaczmarek et al., 2002). Parental cultivars have been selected to achieve a great diversity among the progeny (DH lines) in relation to malting quality characters. Doubled haploid lines, the parental genotypes, F_2 and F_3 Grit × Havila hybrids and the standard cultivar Rudzik were studied in three locations (Cerekwica, Kruszwica, Łagiewniki) over two years. Each year in each locality experiments were carried out with the same genotypes in the randomized complete block design with three replications. Among various malt characters have been measured in these experiments, genetic parameters for coarse extract yield were be estimated and tested with regard to their stability.

3   Specification of the model and statistical analysis

The statistical model for the experiments repeated with the same set of genotypes at several locations over a period of years was described by Caliński et al. (1997). The analysis involves the use of ANOVA and MANOVA techniques for testing various hypotheses, in particular the hypotheses on genotype main effects and on the interactions of genotypes with locations and years (environments).

Assume that I genotypes are compared in a series of N experiments carried out at J locations over a period of K years. Each of the N experiments is carried out in a randomized complete block design with the same number, L, of blocks. Then the model for the average value of observed trait can be written for the vector of genotypes, $\mathbf{y}_{jk}$, in the form

$$\mathbf{y}_{jk} = \mu + \alpha^L(j) - \alpha^T(k) + \mathbf{a}^E(j,k) + \mathbf{e}_{jk}, \tag{1}$$

where $\mathbf{y}_{jk} = [\mathbf{y}_{1jk}, \mathbf{y}_{2jk}, ..., \mathbf{y}_{Ijk}]'$ is the vector of observations of genotypes in location j and year k, $(j = 1, 2, ..., J; k = 1, 2, ..., K)$, $\mu = [\mu_1, \mu_2, ..., \mu_I]'$ is the vector of the fixed average values of genotype i $(= 1, 2, ..., I)$ over all locations and years, $\alpha^L(j) = [\alpha_1^L(j), \alpha_2^L(j), ..., \alpha_I^L(j)]'$, $\alpha^T(k) = [\alpha_1^T(k), \alpha_2^T(k), ..., \alpha_I^T(k)]'$ are the vectors of the fixed location and year effects respectively, $\mathbf{a}^E(j,k) = [a_1^E(j,k), a_2^E(j,k), ..., a_I^E(j,k)]'$ is the vector of random effects $a_i^E(j,k)$ being the deviation of the capacity of genotype i under the environment of the site of the experiment at location j in year k, and $\mathbf{e}_{jk} = [e_{1jk}, e_{2jk}, ..., e_{Ijk}]'$ is the random vector of average errors from the experiments.

Assuming normality for the independently distribution random vector (1) one can write

$$\mathbf{y}_{jk} \sim N_I(\mu + \alpha^L(j) + \alpha^T(k), \mathbf{\Sigma}_y), \tag{2}$$

with the description matrix

$$\mathbf{\Sigma}_y = \mathbf{\Sigma}_m + (\sigma_e^2/L)\mathbf{I}_I, \tag{3}$$

where $\mathbf{\Sigma}_m = [\sigma_{ii'}(j,k)] = [\sigma_{ii'}]$, $(i, i' = 1, 2, ..., I)$, for any j and k denotes the dispersion matrix. No restrictions are imposed on the structure of this matrix, but is assumed that is common for all the environments. As the error, the usual assumptions is made. For the estimation purposes no other assumption are needed.

Now, using the centring matrix $\mathbf{G} = \mathbf{I}_I - \mathbf{I}^{-1}\mathbf{1}_I\mathbf{1}_I'$ it is convenient to transform (1) into the model

$$\mathbf{Z}_{jk} = \mathbf{G}\mathbf{y}_{jk} = \alpha^G + \alpha^{GL}(j) + \alpha^{GT}(k) + \mathbf{a}^{GE}(j,k) + \mathbf{f}_{jk}, \tag{4}$$

where the vector $\alpha^G = \mathbf{G}\mu$ is composed of the genotype main effects, $\alpha^{GL}(j) = \mathbf{G}\alpha^L(j)$ of the genotype interactions with location j, $\alpha^{GT}(k) = \mathbf{G}\alpha^T(k)$ of the genotype interactions with year k, $\mathbf{a}^{GE}(j,k) = \mathbf{G}\mathbf{a}^E(j,k)$ of the genotype interactions with the environment of the site of the experiment at location j in year k, and $\mathbf{f}_{jk} = \mathbf{G}\mathbf{e}_{jk}$ is composed of the genotype error deviations from the average experimental error.

The model (4) allows to estimate the vector of genotype main effects α^G as well the vector of genotype contrasts $\mathbf{c}_p'\alpha^G$ if $\mathbf{c}_p$ is any vector such that $\mathbf{c}_p'\mathbf{1}_I = 0$.

In addition to that the following hypotheses can be tested:
– the hypotheses concerning particular contrasts between genotypes, $H_{c_p'G}$: $\mathbf{c}_p'\alpha^G = 0$, with the Hotelling T^2, statistic and
– the hypotheses of no interactions between the contrast of genotypes and environment $H_{c_p'GE} : var\{\mathbf{c}_p'\alpha^{GE}(j,k) = 0$ for all j and k, with F-statistic.

4 Genetic analysis

The model of observations for series of experiments presented above can be applied to the study of genes effects on the basis of doubled haploid lines and F_2 and F_3 hybrid generations. Interested parameters in this context are additive gene effects $[d]$, dominance effects $[h]$, homozygous $\times$ homozygous interaction effects $[i]$ and heterozygous $\times$ heterozygous interaction effects $[l]$. These parameters can be defined in terms of some linear combinations (contrasts) among the genotype effects (Adamski, 1993). Their estimators in a vector notation are as follows:

$$[\hat{d}] = \mathbf{c}_{[d]}'\hat{\alpha}^G, \quad [\hat{h}] = \mathbf{c}_{[h]}'\hat{\alpha}^G, \quad [\hat{\imath}] = \mathbf{c}_{[i]}'\hat{\alpha}^G, \quad [\hat{l}] = \mathbf{c}_{[l]}'\hat{\alpha}^G,$$

TABLE 1. Estimates of gene effects for coarse extract yield in particular environments

Gene effect	Environment					
	Year 1			Year 2		
	K	L	C	K	L	C
Additive $[d]$	3.97	2.01	3.40	2.95	6.33	2.50
Dominance $[h]$	-7.29	0.44	10.69	11.67	3.61	9.91
Epistasis:						
Homo×homo $[i]$	-0.30	-0.89	-0.42	0.05	3.28	0.01
Hetero×hetero $[l]$	11.35	-1.70	-14.07	-15.38	-5.08	-21.38

where $\hat{\alpha}^G = [\hat{\alpha}(DH_m), \hat{\alpha}(DH_{\max}), \hat{\alpha}(DH_{\min}), \hat{\alpha}(F_2), \hat{\alpha}(F_3)]$ is a vector of the generation main effects of studied traits and $c_{[d]}$, $c_{[h]}$, $c_{[i]}$, $c_{[l]}$ are the vectors of the correspond coefficients between generations such that $c'_{[d]}\mathbf{1} = \mathbf{c}'_{[h]}\mathbf{1} = \mathbf{c}'_{[i]}\mathbf{1} = \mathbf{c}'_{[l]}\mathbf{1} = 0$.
For the data from the experiments with DH lines and F_2, F_3 hybrids the coefficients of contrasts concerning genetic parameters can be written as

$$
\begin{aligned}
c'_{[d]} &= [\quad 0 \quad 0.5 \quad -0.5 \quad 0 \quad 0]', \\
c'_{[h]} &= [\quad -6 \quad 0 \quad 0 \quad -2 \quad 8]', \\
c'_{[i]} &= [\quad -1 \quad 0.5 \quad -0.5 \quad 0 \quad 0]', \\
c'_{[l]} &= [\quad 8 \quad 0 \quad 0 \quad 8 \quad -16]'.
\end{aligned}
$$

5   Analysis of the data

Statistical calculation of the data described in Section 2 were made by the computer program SERGEN (Caliński et al., 1998). Observed traits was of normal distribution. Estimates of genetic parameters for coarse extract yield were found for each of the six environments (Table 1). Mean estimates of gene effects over environments and results of testing of their significance are presented in Table 2.

Analysis of coarse extract yield indicates that additive effects estimated over environments were significant. Mean estimates of the other gene effects were not significant. Interaction of additive effects and homozygous × homozygous epistasis effects with environments was very high, whereas the dominance effects and heterozygous × heterozygous epistasis effects were stable.

...

TABLE 2. Mean estimates of gene effects for coarse extract yield and results of testing the hypotheses concerning their interaction with environments

Gene effect	Estimate	F-statistic value for	
		gene effect	interaction
Additive [d]	3.53	31.69	20.48
Dominance [h]	4.84	2.56	2.20
Epistasis:			
Homo × homo [i]	0.29	0.22	6.59
Hetero × hetero [l]	-7.71	2.58	1.57
Critical values:			
$F_{0.05}$		6.61	2.24
$F_{0.01}$		16.26	3.06

References

Adamski T. (1993). The use of doubled haploid lines for genetic analysis of quantitative traits (in Polish). *Treatises and Monographs* Nr. 2, IGR PAN, Poznań.

Caliński T., Czajka S., Kaczmarek Z. (1997). A multivariate approach to analysing genotype-environment interaction. In: *Advances in Biometrical Genetic* (P. Krajewski and Z. Kaczmarek, eds.), Poznań, 3-14.

Caliński T., Czajka S., Kaczmarek Z. , Krajewski P., Siatkowski I. (1998). *Sergen 3 - User's Guide. Statistical methodology and usage of the program SERGEN dedicated to "Analysis of series of plant genetic and breeding experiments"*. IGR PAN, Poznań.

Kaczmarek Z., Surma M., Adamski T., Jeżowski S., Madajewski R., Krystkowiak K., Kuczyńska A. (2002). Interaction of gene effects with environments for malting quality of barley doubled haploids. *J. Applied Genetics*, 43(1), 33-42.

Mather K., Jinks J.L. (1982). *Biometrical Genetics* (3rd edn.). Chapman and Hall, London.

Surma M., Kaczmarek Z., Adamski T. (1997). Estimation of genetic parameters based on doubled haploids and early generation. In: *Advances in Biometrical Genetics.* (P. Krajewski and Z. Kaczmarek, eds.), Poznań, 281-284.

Split-plot × Split-block type three factor designs

Katarzyna Ambroży[1], Iwona Mejza[1]

[1] Department of Mathematical and Statistical Methods, Agricultural University, Wojska Polskiego 28, 60-637 Poznań, Poland.

Abstract: We consider modelling and some construction methods of incomplete split-plot × split-block designs for three factor experiments. In the modelling we take into account a structure of an experimental material and a four-step randomization schema. We adopt the approach typical to multistratum experiments with orthogonal block structure with respect to the analysis of the obtained randomization model with seven strata. A brief discussion connected with the method of the construction of the design is given.

1   Introduction

The purpose of this paper is to present a method of designing three factor experiments and modelling data obtained from them. We are interested in one of so called mixed designs combined of a split-plot design and a split-block design (e.g. Gomez and Gomez, 1984). Another mixed design of a split-block-plot type was presented in the paper by Mejza I. and Ambroży (2003). That design was an extension of a split-block design in which each intersection plot was divided into subplots to accommodate a third factor. So the third factor was in a split-plot design in a relation to row and column treatments (i.e. combinations of levels of the two first factors).

In this paper we present another arrangement of units in the three factor designs. In field experiments certain treatments such as types of cultivation, application of irrigation water etc., may be necessary to be arranged in strips (rows or columns) across each block. Then it is convenient to arrange the plots of the design in the following way: the columns (or the rows) of the split-block design are split into smaller strips to accommodate the third factor. So, the third factor will be in the split-plot design in a relation to the column (or row) treatments. The new design obtained this way will be called the split-plot × split-block (shortly SPSB) design. We will consider incomplete (in particular complete) SPSB designs (i.e. when a number of the levels of at least one factor is larger or equal than the number of appropriate for them strips within each block).

2   Assumptions and notations

Let us consider a three-factor experiment of a SPSB type in which the first factor, say A, has s levels A_1, A_2, ..., A_s, the second factor, say B, has t levels B_1, B_2, ..., B_t and the third factor, say C, has w levels C_1, C_2, ...,C_w. Thus the number $v = stw$ denotes the number of all treatment combinations in the experiment. The experimental material is assumed to be divided into b blocks each of a row-column structure with k_1 rows $(k_1 \leq s)$ and k_2 columns of the first order, shortly, columns I $(k_2 \leq t)$. So within each block there are k_1k_2 intersection plots of the first order called whole plots. Then each column I has to be split into k_3 columns of the second order, shortly, columns II $(k_3 \leq w)$. So there are $k_1k_2k_3$ intersection plots of the second order called small plots within each block. Here the rows correspond to the levels of the factor A (row treatments), the columns I correspond to the levels of the factor B (column I treatments), and the columns II are to accommodate the levels of the factor C (column II treatments). The order of the arrangement of the factors in the designs considered is very important from the statistical point of view. This affects the precision of contrasts estimation concerning main effects and interaction effects of the factors.

3   Linear model and its analysis

We consider a randomization model of observations, in which a form and properties are strictly connected with the performed randomization processes in the experiment. The randomization scheme used here consists of four randomization steps performed independently, i.e. by randomly permuting blocks within total experimental material, by randomly permuting rows within the blocks, by randomly permuting columns I within blocks and by randomly permuting columns II within the column I in each block. Three of the randomization processes proceed as in a split-block design and refer to the blocks, the rows and the columns I. The fourth step, relating to the columns II, is performed as in a split-plot design. It is worth noticing that one can start the randomization scheme conversely, i.e. first performing three randomizations as in the split-plot design (the blocks, the columns I and the column II) and then the fourth step as in the split-block design (the rows). The ordering of these processes does not matter for the form of the obtained by this way model of observations. Then, assuming the usual unit-treatment additivity and uncorrelation of the technical errors, with zero expectation and a constant variance σ_e^2, the model can be written as

$$\mathbf{y} = \mathbf{\Delta}'\tau + \sum_{f=1}^{6} \mathbf{D}_f'\xi_f + \mathbf{e} \tag{1}$$

where $\mathbf{y}$ is a dimensional vector of lexicographically ordered observations, $\boldsymbol{\Delta}'$ $(n \times v)$ is a known design matrix for v treatment combinations, $n = bk_1k_2k_3$, $\mathbf{D}_1'$ $(n \times b)$, $\mathbf{D}_2'$ $(n \times bk_1)$, $\mathbf{D}_3'$ $(n \times bk_2)$, $\mathbf{D}_4'$ $(n \times bk_2k_3)$, $\mathbf{D}_5'$ $(n \times bk_1k_2)$, $\mathbf{D}_6'$ $(n \times n)$ are design matrices for blocks, rows (within blocks), columns I (within blocks), column II (within columns I), whole plots (within blocks) and subplots (within whole plots) respectively, τ $(v \times 1)$ is the vector of fixed treatment combination effects, ξ_1 $(b \times 1)$, ξ_2 $(bk_1 \times 1)$, ξ_3 $(bk_2 \times 1)$, ξ_4 $(bk_2k_3 \times 1)$, ξ_5 $(bk_1k_2 \times 1)$, ξ_6 $(bk_1k_2k_3 \times 1)$, $\mathbf{e}$ $(n \times 1)$ are random effect vectors of blocks, rows, columns I, columns II, whole plots, subplots and technical errors, respectively.

Let σ_f^2 $(f = 1, 2, ..., 6)$ denote, respectively, the variances of the effects of the blocks, the rows, the columns I, the columns II, the whole plots, the subplots. Then under our assumptions we can write the first two moments of distributions of the random variables ξ_f $(f = 1, 2, ..., 6)$, i.e. $\mathrm{E}(\xi_f) = \mathbf{0}$, $\mathrm{Cov}(\xi_f, \xi_{f'}) = \mathbf{V}_f$, for all $f = f'$ and $\mathrm{Cov}(\xi_f, \xi_{f'}) = \mathbf{0}$, for all $f \neq f'$. Thus the considered dispersion structure of the linear model has the form

$$\mathrm{Cov}\,(\mathbf{y}) = \sum_{f=1}^{6} \mathbf{D}_f' \mathbf{V}_f \mathbf{D}_f + \sigma^2 \mathbf{I}_n \tag{2}$$

It is easy to show (cf. Ambroży and Mejza I., 2003) that the dispersion matrix (2) can be written as $\mathrm{Cov}(\mathbf{y}) = \sum_{f=0}^{6} \gamma_f \mathbf{P}_f$, where $\gamma_0 = \sigma_e^2$, $\gamma_1 = k_1k_2k_3\sigma_1^2 + \sigma_e^2$, $\gamma_2 = k_2k_3\sigma_2^2 + \sigma_e^2$, $\gamma_3 = k_1k_3\sigma_3^2 + \sigma_e^2$, $\gamma_4 = k_1\sigma_4^2 + \sigma_e^2$, $\gamma_5 = k_3\sigma_5^2 + \sigma_e^2$, $\gamma_6 = \sigma_6^2 + \sigma_e^2$ and $\{\mathbf{P}_f\}$, $f = 0, 1, ..., 6$, are a set of pairwise orthogonal matrices summing to the identity matrix. The range space of $\mathbf{P}_f$ is termed the f-th stratum with $\mathbf{P}_f$ being orthogonal projection onto this stratum. It follows that the considered design has an orthogonal block structure (cf. Nelder, 1965, Houtman and Speed, 1983). So the model can be analysed using the methods developed for multistratum experiments. In this case, we have zero stratum (0) generated by the vector of ones, inter-block stratum (1), inter-row (within the block) stratum (2), inter-column I (within the block) stratum (3), inter-column II stratum (4) (within the column I), inter-whole plot (within the block) stratum (5), and inter-subplot (within the whole plot) stratum (6). The statistical analysis of such model is connected with the algebraic properties of stratum information matrices for the treatment combinations in the incomplete SPSB designs $\mathbf{A}_f$, $f = 0, 1, ..., 6$ (cf. Ambroży and Mejza I, 2003). The obtained designs will be characterized with respect to (shortly w.r.t.) the general balance property and stratum efficiency factors of the design for a set of orthogonal contrasts between the treatment combination effects. These efficiency factors are eigenvalues of the information matrices $\mathbf{A}_f$, $f = 1, 2, ..6$ w.r.t. $\mathbf{r}^\delta$, where $\mathbf{r}$ is the vector of replications of the treatment combinations and $\mathbf{r}^\delta = \mathrm{diag}(r_1, r_2, ..., r_v)$. The contrasts are connected with the comparisons

among the main effects of the considered factors and the interaction effects
between them.

4   Construction method of SPSB type designs

We will introduce abbreviations to describe the properties such as efficiency and balance of the design. Let $M_f\{q,\alpha\}$ denote the property that q contrasts among the treatments of factor M (or interaction contrasts) are estimated with the efficiency α in the f-th stratum. In other words, we say that the design is $M_f\{q,\alpha\}$ - balanced or $M_f\{q,1\}$ - orthogonal.

Let $\mathbf{N}_A(s\times b)$, $\mathbf{N}_B(t\times b)$ and $\mathbf{N}_C(w\times b)$ be incidence matrices of subdesigns for the row treatments, the column I treatments and the column II with respect to the blocks, respectively. In the present paper the construction method for three factor experiments is based on Kronecker product of matrices denoted by $\otimes$. Then we have $\mathbf{N}_1 = \mathbf{N}_A \otimes \mathbf{N}_B \otimes \mathbf{N}_C$, where $\mathbf{N}_1$ is the treatment combinations vs. blocks incidence matrix of the SPSB design. Let

$\mathbf{C}_A = \mathbf{r}_A^\delta - k_1^{-1}\mathbf{N}_A\mathbf{N}_A'$ with nonzero eigenvalues $\mu_1,\mu_2,...,\mu_{s-1}$ w.r.t. $\mathbf{r}_A^\delta$,
$\mathbf{C}_B = \mathbf{r}_B^\delta - k_2^{-1}\mathbf{N}_B\mathbf{N}_B'$ with nonzero eigenvalues $\xi_1,\xi_2,...,\xi_{t-1}$ w.r.t. $\mathbf{r}_B^\delta$,
$\mathbf{C}_C = \mathbf{r}_C^\delta - k_3^{-1}\mathbf{N}_C\mathbf{N}_C'$ with nonzero eigenvalues $\psi_1,\psi_2,...,\psi_{w-1}$ w.r.t. $\mathbf{r}_C^\delta$

be the information matrices for the treatments of the factors A, B and C, respectively, in the subdesigns.

Following algebraic properties of the information matrices of the SPSB design and the subdesigns we have:

Corollary. The incomplete SPSB design based on Kronecker product of matrices is:

$A_1\{1, 1-\mu_h\}$ - balanced and $A_2\{1,\mu_h\}$ - balanced, $h = 1,2,...,s-1$,
$B_1\{1, 1-\xi_m\}$ - balanced and $B_3\{1,\xi_m\}$ - balanced, $m = 1,2,...,t-1$,
$C_1\{1, 1-\psi_g\}$ - balanced and $C_4\{1,\psi_g\}$ - balanced, $g = 1,2,...,w-1$,
$(A\times B)_1\{1, (1-\mu_h)(1-\xi_m)\}$ - balanced, $(A\times B)_2\{1,\mu_h(1-\xi_m)\}$ - balanced, $(A\times B)_3\{1, (1-\mu_h)\xi_m\}$ - balanced and $(A\times B)_5\{1,\mu_h\xi_m\}$ - balanced, $h = 1,2,...,s-1$, $m = 1,2,...,t-1$,
$(A\times C)_1\{1, (1-\mu_h)(1-\psi_g)\}$ - balanced, $(A\times C)_2\{1, \mu_h(1-\psi_g)\}$ - balanced, $(A\times C)_4\{1, (1-\mu_h)\psi_g\}$ - balanced and $(A\times C)_6\{1, \mu_h\psi_g\}$ - balanced, $h = 1,2,...,s-1$, $g = 1,2,...,w-1$,
$(B\times C)_1\{1, (1-\xi_m)(1-\psi_g)\}$ - balanced, $(B\times C)_3\{1, \xi_m(1-\psi_g)\}$ - balanced, $(B\times C)_4\{1,\psi_g\}$ - balanced, $m = 1,2,...,t-1$, $g = 1,2,...,w-1$,
$(A\times B\times C)_1\{1, (1-\mu_h)(1-\xi_m)(1-\psi_g)\}$ - balanced, $(A\times B\times C)_2\{1, \mu_h(1-\xi_m)(1-\psi_g)\}$ - balanced, $(A\times B\times C)_3\{1, (1-\mu_h)\xi_m(1-\psi_g)\}$ - balanced, $(A\times B\times C)_4\{1, (1-\mu_h)\psi_g\}$ - balanced, $(A\times B\times C)_5\{1, \mu_h\xi_m(1-\psi_g)\}$ - balanced, $(A\times B\times C)_6\{1, \mu_h\psi_g\}$ - balanced, $h = 1,2,...,s-1$, $m = 1,2,...,t-1$, $g = 1,2,...,w-1$.

We can notice that all contrasts connected with main effects of the factors are estimable at most in two strata only (the inter-block stratum and the

appropriate stratum for each factor). The interaction contrasts (between the combination effects of two factors) can be estimable at most in four different strata and others (between the combination effects of three factors) at most in all strata. In any case the number of efficiency balanced classes for the same type of the contrasts will suffer reduction, when at least one of the subdesigns will be efficiency balanced (or orthogonal) block designs (cf. Caliński and Kageyama, 1996).

As an example, let us consider a $2 \times 3 \times 4$ - factorial experiment in order to determine an effect of irrigation, nitrogen fertilization and chemical protection on winter wheat disease infestation. An experimental material was limited, hence the experiment was carried out in incomplete SPSB design according to the incidence matrix $\mathbf{N}_1 = \mathbf{1}_2 \otimes \mathbf{1}_3 \otimes \mathbf{N}_C$, where $\mathbf{N}_C$ is the incidence matrix of BIB design with blocks (1, 2) (3, 4) (1, 3) (2, 4) (1, 4) (2, 3). The eigenvalues of the matrix $\mathbf{C}_C$ are equal to $\psi_1 = \psi_2 = \psi_3 = 2/3$ w.r.t. $\mathbf{r}_C^\delta = 3\mathbf{I}_4$. Finally, the parameters of the SPSB design were: $v = 24$, $k_1 = s = 2$, $k_2 = t = 3$, $k_3 = 2$, $w = 4$, $b = 6$ and the efficiency of the SPSB design w.r.t. the comparisons among the main effects and the interaction effects was following:

$A_2\{1, 1\}$ - orthogonal, $B_3\{2, 1\}$ - orthogonal,
$C_1\{3, 1/3\}$- balanced and $C_4\{3, 2/3\}$ - balanced,
$(A \times B)_5\{2, 1\}$ - orthogonal,
$(A \times C)_2\{3, 1/3\}$ - balanced, $(A \times C)_6\{3, 2/3\}$ - balanced,
$(B \times C)_3\{6, 1/3\}$ - balanced and $(B \times C)_4\{6, 2/3\}$ - balanced,
$(A \times B \times C)_5\{6, 1/3\}$ - balanced, $(A \times B \times C)_6\{6, 2/3\}$ - balanced.

References

Ambroży, K., Mejza, I. (2003). Some split-plot × split-block designs. *Colloq. Biom.*, **33**, 83-96.

Caliński, T., Kageyama, S. (1996). *Block designs: their combinatorial and statistical properties*. S. Ghosh and C.R. Rao, eds., Handbook of Statistics, Vol. 13, 809-873.

Gomez, K.A., Gomez, A.A. (1984). *Statistical procedures for agricultural research*. Wiley, New York.

Houtman, A.M., Speed, T.P. (1983). Balance in designed experiments with orthogonal block structure. *Ann. Statist.*, **11**, 1069-1085.

Mejza, I., Ambroży, K. (2003). Modelling some experiments carried out in incomplete split-block-plot designs. *Proc. of the 18th IWSM*, 293-297.

Nelder, J.A. (1965). The analysis of randomized experiments with orthogonal block structure. *Proc. Roy. Soc. Lond.*, Ser. A, **283**, 147-178.

Optimization of Fiber Tracking in Human Brain Mapping: Statistical Challenges

Heim S.[1,2], Hahn K.[3], Auer D. P.[2,4] and Fahrmeir L.[1]

[1] Institute of Statistics, Ludwig–Maximilians–University, Munich, Germany
[2] NMR Research Group, Max–Planck–Institute of Psychiatry, Munich, Germany
[3] Institute of Biomathematics and Biometry, GSF, Neuherberg, Germany
[4] Academic Radiology, University of Nottingham, England

Abstract: The principle of fiber tracking on the basis of diffusion tensor imaging is explained. Challenges in developing new statistical methods or in adapting well-known techniques are pointed out.

1 Background

During the last decade of neuroscience, diffusion magnetic resonance imaging (DTI) has become a powerful tool for the quantification of ultrastructural tissue properties which is of prime interest for monitoring major diseases such as acute ischaemia and multiple sclerosis. A second important benefit is to non–invasively determine fiber tracts which may be of impact for neurosurgical planning. Thus allowing the identification of anatomical connections between different brain regions, DTI supplements the visualization of functional brain areas by functional magnetic resonance imaging. The biophysical basis of DTI is the random diffusion of water molecules which depends on the surrounding tissue structure and can mathematically be conceptualized by a 3d Brownian process with location dependent diffusion matrix $D(x_t)$ at x_t :

$$dx_t = D^{\frac{1}{2}}(x_t)dw_t, \tag{1}$$

where $t \geq 0$ is "time" after starting from a seed point x_0, and w_t is a 3d standard Wiener process. As cerebral white matter is highly organized in the ultrastructural level, random motion of particles preferentially follows the direction of densely packed fiber bundles. This phenomenon ('anisotropy') is captured in the so–called tensor model, i. e. the symmetric positive definite (3×3)–diffusion matrix $D(x_t)$. Diagonalization provides eigenvectors which correspond to the principal orthogonal diffusion directions, whereas the respective eigenvalues reflect the diffusion strength along
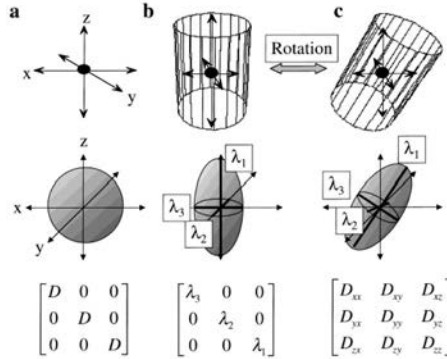
FIGURE 1. Geometrical interpretation of the diffusion tensor.

each axis (Fig. 1). Exploiting this information, the tensor model allows to identify neuronal fibers (see Basser et al. 2002 for a review).

2   Data Basis

Concerning the available datasets, diffusion weighted images are recorded in six non-collinear directions on a 1.5 T human scanner with a resulting image matrix of $128 \times 128 \times 24$ at a resolution of $18.75 \times 1.875 \times 4$ mm^3. For each voxel v, the six free tensor parameters $d(v) = (D_{xx}, D_{xy}, D_{xz}, D_{yy}, D_{yz}, D_{zz})$ are estimated from the logarithmized Stejskal-Tanner equation:

$$ln\left(\frac{S_i(v)}{S_0(v)}\right) = -z_i'd(v) + \varepsilon, i = 1, \ldots, K, \qquad (2)$$

where S_i denotes the signal intensities of the (at least) $K = 6$ diffusion gradient weighted images and S_0 refers to the unweighted reference image; z_i comprises all relevant parameters of the acquisition scheme.

A more reliable estimate of the tensor is gained by collecting repeated measurements (presently three repeats) or considerably enhancing the overall number of encoding directions (Jones et al. 2004). The resulting spatial tensor field represents the data basis for a tracking algorithm. In addition, diverse rotation invariant scalars are derived from the tensor which mainly serve diagnosis and inference of disease stages.

3   Tracking using state space models

While most current line propagation algorithms work deterministically (Mori et al. 2002), Gössl et al. (2002) embedded a discretized version of
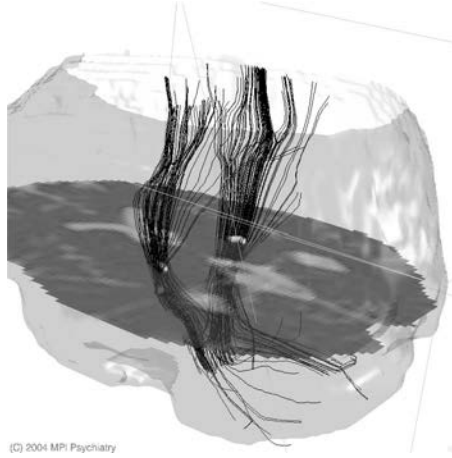
FIGURE 2. Pyramidal tract with superimposed starting regions (white blobs). These fiber bundles represent a major pathway between the motor cortex and spinal cord. A representative slice of the mean diffusivity map has been added for orientation with the light parts belonging to the lateral ventricles.

the Brownian process (Eq. (1))

$$x_t = x_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, D(x_{t-1})) \tag{3}$$

as transition equation for the latent curve x_t in a linear state space model with noised observations:

$$y_t = x_t + \eta_t, \quad \eta_t \sim N(0, \sigma^2 I). \tag{4}$$

In contrast to a conventional linear state space model, $y_t, t = 1, \ldots, T$ have to be generated from the diffusion tensor data acquired as in Section 2. The noisy (pseudo–) observations y_t of x_t can be sequentially obtained from

$$y_t = \hat{x}_{t-1} + ev_{t-1}, \quad t = 1, 2, \ldots \tag{5}$$

with $\hat{x}_{t-1}$ estimated current state of curve and ev_{t-1} principal eigenvector of the tensor $D(\hat{x}_{t-1})$. A step size parameter and a constraint for avoiding too wiggly and unplausible, highly curved fibers are additionally introduced. Therefore, recursive application of the Kalman filter and smoother provides fairly smooth estimates of trajectories (Fig. 2).

4   Problems

DTI is prone to numerous detrimental sources of artefacts which may impair data reliability and validity (Basser et al. 2002, Mori et al. 2002). A

major consequential problem is the uncontrolled prolongation of such arte-
facts into derived parameters causing both random and systematic errors.
In particular, uncertainties in the principal eigenvector may lead to erro-
neous 3d fiber reconstruction. Furthermore, voxels can occur with a more
disc-shaped tensor containing ambiguous geometrical information: among
diverse conditions, it may indicate a voxel of other tissue than white mat-
ter, a voxel contaminated by a second tissue type (partial volume effect)
or a voxel containing crossing fiber bundles.

5   Approaches and Statistical Challenges

In order to improve data quality, data preprocessing focuses on correcting
the measured signal intensities or the derived tensors. Hahn et al. (2004)
recently implemented a sophisticated edge preserving smoothing algorithm
which also proves superior for DTI data in comparison with the more widely
applied Gaussian filter that may result in undesirably blurred data.

Also tackling the problem of data reliability, we generated an objective
quality rating for real raw data using nonparametric bootstrapping and
investigated its sensitivity to a selection of intrinsic and extraneous in-
fluencing factors (Heim et al. 2003). In brief, $N = 100$ resamples were
obtained for each individual dataset by drawing with replacement from
the corresponding repeated measurements of each applied gradient direc-
tion. The respective N tensor maps provided N maps of scalar measures
of the anisotropy and voxelwise bootstrap estimates of confidence intervals
as well as coefficients of variation of these measures. Appropriate aggre-
gation within areas of interest yielded global measures for quantifying the
statistical uncertainty of scalar measures and its additional dependence on
different tissue types.

While the uncertainty of the principal diffusion direction, i. e. the main
eigenvector of the tensor, has been explored on a single voxel level (Jones
et al. 2003), evaluating the regional and global uncertainty of tracking
results is still to be realized.

So far, the preferably denoised tensor is independently estimated based on
the linear regression model (Eq. (2)) for each voxel v. A more complex ten-
sor estimation could take into account spatial correlation and information
from neighboring voxels. For this purpose, the location dependent tensor
elements $d(v)$ in Eq. (2) are treated as space-varying regression coefficients,
each of which can be nonparametrically approximated by a linear combi-
nation of basis functions $B_j(v)$, e. g. tensor product splines or radial basis
functions:

$$D_l(v) = \sum \beta_j B_{j,l}(v), \quad l = xx, xy, \ldots, zz \tag{6}$$

Spatial smoothing can be introduced by appropriate spatial penalties for
the coefficients β_j of neighboring voxels.

This 3d surface smoothing of the tensor elements yields an effective refinement of the underlying data grid since it allows to estimate the diffusion tensor at each arbitrary position. Hence, a more reliable and precise tracking is enabled, especially when ambiguity is caused by partial volume effects due to the coarse spatial resolution compared with the size of uniform fiber tracts.

Concerning the issue of fiber crossing, the possibly available information of the associated fiber ending has not been exploited so far. We plan to incorporate the end point information into the existing algorithm within the framework of a Brownian bridge to further improve the tracking results.

References

Basser P.J., Jones D.K. (2002). Diffusion-tensor MRI: theory, experimental design and data analysis – a technical review. *NMR Biomed*, **15**, 456-467.

Gössl C., Fahrmeir L., Pütz B., Auer L.M., Auer D.P. (2002). Fiber tracking from DTI using linear state space models: detectability of the pyramidal tract. *Neuro Image*, **16**, 378-388.

Hahn K.R., Prigarin S., Pütz B., Hasan K.M. (2004). Spatial smoothing for diffusion tensor imaging with low signal to noise ratios. Submitted to *Neuro Image*.

Heim S., Hahn K., Sämann P.G., Fahrmeir L., Auer D.P. (2004). On the assessment of DTI data quality using bootstrap analysis. To appear in *Magn Reson Med*.

Jones D.K. (2003). Determining and visualizing uncertainty in estimates of fiber orientation from diffusion tensor MRI. *Magn Reson Med*, **49**, 7-12.

Jones D.K. (2004). The effect of gradient sampling schemes on measures derived from diffusion tensor MRI: a Monte Carlo study. *Magn Reson Med*, **51**, 807-815.

Mori S., van Zijl P.C. (2002). Fiber tracking: principles and strategies - a technical review. *NMR Biomed*, **15**, 468-480.

Estimates of the short term effects of air pollution in Italy using alternative modelling techniques

Michela Baccini[1], Annibale Biggeri[1], Gabriele Accetta[2], Corrado Lagazio[3], Aitana Lertxundi[1] and Joel Schwartz[4]

[1] Department of Statistics "G. Parenti", University of Florence, Italy
[2] Department of Epidemiology - ASL ROMA E, Italy
[3] Department of Statistics, University of Udine, Italy
[4] Department of Environmental Health, Harvard School of Public Health, Boston

Abstract: In the analysis of short term effect of air pollution on health, methods able to control for nonlinear confounding effect of temporal trend are required. We analyze the association between PM10 daily concentrations and Mortality/Hospital Admissions in the Italian Meta-analysis of Short-term effects of Air pollutants (MISA), using alternative modeling techniques: Generalized Additive Models with penalized regression spline fitted by the direct method in R software (GAM-R) and Generalized Linear Models with natural cubic spline (GLM+NS). We find that the two approaches provide similar results. If we are interested in overall estimates and a random effects meta-analysis model is specified, a certain robustness of results to change number of degrees of freedom for the spline is to be expected.

Keywords: Generalized Additive Model; penalized regression spline; cubic regression spline; epidemiological time series.

1 Introduction

In the analysis of short term effect of air pollution on health, the characteristics of epidemiological time series data require statistical methods able to control for nonlinear confounding effect of temporal trend. In the literature, most of the studies used flexible semi-parametric approaches, specifying Generalized Additive Models (GAMs) with smoothing splines or locally weighted regressions in moving ranges of the data. Recently major concern was raised about numerical accuracy of the estimates of pollutant effect obtained from this kind of models using commercial statistical software which implements backfitting algorithm, namely Splus. Two important critical points were addressed: the *gam* function of Splus provides an approximation of the variance-covariance matrix which takes into account only the linear component of the smooth function, bringing to underestimated standard error for the air pollution effect (Ramsay *et al.*, 2003); this

function uses too bland convergence criteria for the estimation algorithm, producing biased point estimates, whenever the magnitude of the effect to be estimated is small and convergence of backfitting is slow due to relevant amount of concurvity in data (Dominici *et al.*, 2002).

The present paper analyzes data of the Italian Meta-analysis of Short-term Effects of Air Pollution (MISA), using alternative modeling approaches: GLM with natural cubic spline for seasonality (GLM+NS) and GAM with penalized regression spline fitted by the *gam* function of R software (Wood, 2000) (GAM-R). Both these approaches estimate the variance-covariance matrix correctly and are less sensitive to the definition of convergence criteria.

2   Methods

The MISA study investigated the short term effect of air pollution on mortality and hospital admissions in height Italian cities. The analysis was age-adjusted. We controlled for time-related confounding including in the model spline terms, whit pre-defined number of degrees of freedom. Two linear terms constrained to joint in 21 C for temperature and linear and quadratic terms for relative humidity were defined. We controlled for day of the week, holidays and influenza epidemics by appropriate dummy variables (Biggeri *et al.*, 2001).

We produced air pollution effect estimates both under the parametric approach based on GLM+NS and under the semi-parametric approach based on GAM-R. Once the number and position of knots has been defined (knots were placed evenly throughout the covariate values), maximum likelihood estimates of the coefficients of GLM+NS were obtained using standard IRLS algorithms. Effect estimates under GAM-R were obtained using the *gam* function of R, which maximizes the penalized likelihood by a direct method which avoids the iterative process nested in the backfitting algorithm. We fit also GAM with smoothing cubic splines by the *gam* function of Splus with default ($< 10^{-3}$) and stringent ($< 10^{-14}$) convergence criteria (GAM-S), despite this approach is affected by the previously described drawbacks.

The combined meta-analytic estimates were calculated using fixed and random effects models. A sensitivity analysis to change degrees of freedom for the splines in GLM+NS and in GAM-R was conducted. Finally, the impact of non parametric modeling of temperature on pollutant effect estimates was evaluated. In particular we compared the model proposed in MISA with a model where a penalized regression spline for temperature with 7 degrees of freedom was introduced.
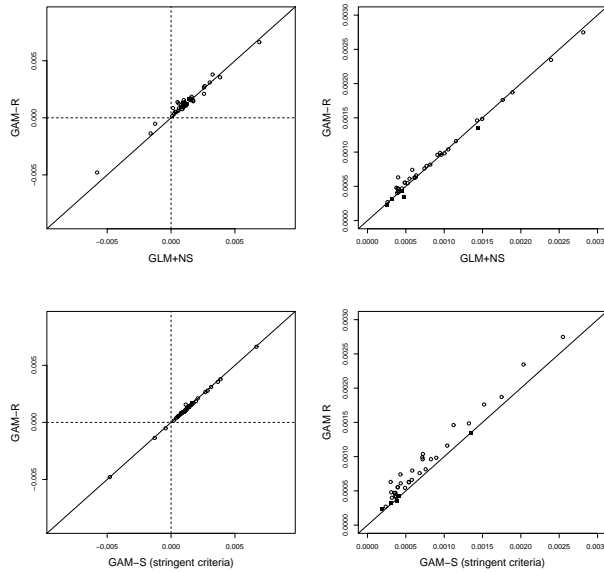
FIGURE 1. MISA 1995-1999. Comparison of city-specific and meta-analytic (in square bold) results for PM10 under different modeling approaches (effect estimates on the left and related standard error estimates on the right).

3   Results

The GLM+NS coefficients estimates resulted generally lower and the estimated standard errors resulted greater, proportionally to their magnitude, than those obtained from GAM-S with default convergence criteria (not reported). Using more stringent convergence criteria, GAM-S provided point estimates very close to those obtained from GAM-R. This is an expected results, when a large number of knots (here 150) is defined for the penalized regression splines. However even if appropriate convergence criteria were defined, performance of GAM-S in terms of estimated precisions did not improve (Fig.1). Results from GAM-R with GLM+NS appeared similar, even if point estimates from GLM+NS resulted usually lower than those obtained from GAM-R.

Addressing attention to meta-analysis results, we can notice that GAM-S with default convergence criteria bringed to overestimated effects and mistakenly small confidence intervals. The overall estimates under GAM-R resulted always slightly higher than under GLM+NS (Table 1 reports results for total mortality).

Overall meta-analytic estimates appeared robust to increasing the number of degrees of freedom for the seasonality splines, both under GAM-R and

TABLE 1. MISA 1995-1999. Combined meta-analytic estimates of percentage increase in total mortality (95% CI) associated to a PM10 increase of 10 $\mu g/m^3$ by fixed and random effects models.

Method	fixed	random
GAM-S default	1.12	1.24
	0.82;1.42	0.63;1.86
GAM-S stringent	0.92	1.06
	0.62;1.22	0.46;1.66
GAM-R	0.90	1.04
	0.55;1.25	0.41;1.67
GLM+NS	0.85	0.98
	0.52;1.18	0.35;1.61

GLM+NS (Figure 2 reports results for total mortality). On the contrary, as the number of degrees of freedom decreased, higher point overall estimates were obtained. This behavior was more evident for GAM-R and if fixed effects meta-analysis was used. Due to the precision of city-specific estimates usually decreased as the number of degrees of freedom increased (not reported), the coefficient of variation calculated under the fixed effects model uniformly increased, the confidence interval for the PM10 effect obtained using 3 degrees of freedom resulting the narrowest. Combining the city-specific results by random effects meta-analysis, a different behavior was observed. The estimated variance decreased then increased, with minimum around 5 degrees of freedom per year (our choice in MISA). When few degrees of freedom for the spline were used, the lower within city variance estimates were balanced by a larger among cities variability.

Results appeared robust to changing the modeling strategies for temperature both in terms of point estimates and precision (not reported).

4   Discussion

In the context of epidemiological time series, using GAM-S can bring to bad city-specific inference and should be avoided. GLM+NS and GAM-R give close results both in city-specific analysis and in meta-analysis. The small observed discrepancy between point estimation under the two approaches can be explained looking at the asymptotic properties of the two methods (Rice, 1986).

When the the random effect meta-analysis model is used, overall point estimates did not appear much sensitive to changing number of degrees of freedom for the spline both under GAM-R and GLM+NS, however a trade-off between overall effect and variance is observed.

The strategy adopted to adjust for the confounding effect of temperature did not appear a major problem.

Fixed Effects Meta-analysis

&

Random Effects Meta-analysis

&
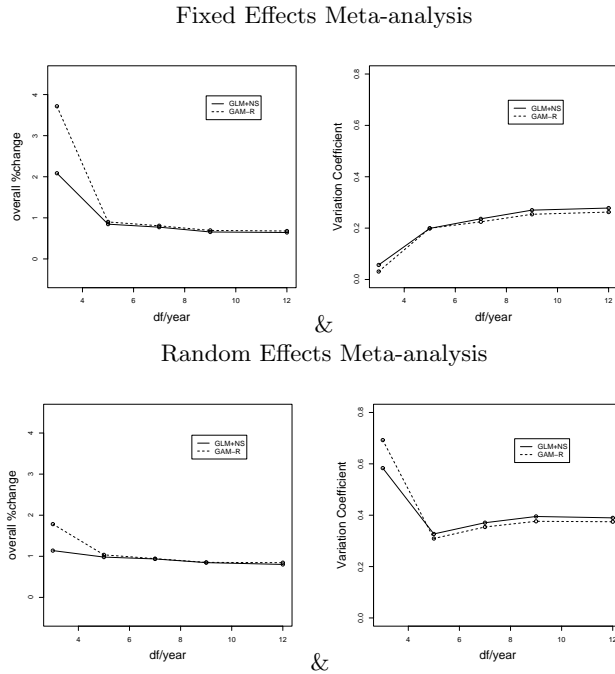
FIGURE 2. MISA 1995-1999. Meta-analysis results for the effect of PM10 on total mortality under GLM+NS and GAM-R, varying the number of degrees of freedom for the seasonality spline.

References

Biggeri, A., Bellini, P. and Terracini, B. (2001) Meta-analysis of the italian studies on short-term effects of air pollution. *Epidemiologia e Prevenzione*, **25**(suppl), 1-72. (italian)

Dominici, F., McDermott, A., Zeger, S.L. and Samet, J. (2002) On the use of Generalized Additive Models in Time-series Studies of Air Pollution and Health. *American Journal of Epidemiology*, **156**, 193-203.

Ramsay, T.O., Burnett, R.T. and Krewski, D. (2003) The Effects of Concurvity in Generalized Additive Models Linking Mortality to Ambient Particulate Matter. *Epidemiology*, **14**, 18-23.

Rice, J. (1986) Convergence rates for partially splined models. *Statist. Probabil. Letters*, **4**, 203-208.

Wood, S.N. (2000) Modelling and Smoothing Parameter Estimation with Multiple Quadratic Penalties. *Journal of the Royal Statistical Society B*, **62**, 413-428.

A multivariate latent Markov model for the analysis of criminal trajectories

Francesco Bartolucci,[1] Fulvia Pennoni[2]

[1] Istituto di Scienze Economiche, Università di Urbino "Carlo Bo", Via Saffi, 42, 61029 Urbino, Francesco.Bartolucci@uniurb.it
[2] Dipartimento di Statistica "G. Parenti", Università di Firenze, Viale Morgagni, 59, 50134 Firenze, pennoni@ds.unifi.it

Abstract: We introduce a multivariate version of the latent Markov model for the investigation of criminal trajectories whose transition matrix may be suitably constrained in order to formulate hypotheses of interest on the criminal behaviour. For the maximum likelihood estimation of the model and its constrained versions we outline an EM-type algorithm. We also illustrate a simple procedure based on the likelihood ratio for choosing the number of states and testing restrictions on the transition matrix.

Keywords: EM algorithm; Latent class model; Hidden Markov processes.

1 Introduction

An important issue in criminology is the analysis of criminal trajectories of a fixed birth cohort followed up for a long period. Among the statistical models that have been used for this kind of analysis (see Francis *et al.*, 2004, and the references therein), the latent Markov model (Wiggins, 1973) seems particularly interesting (Bijleveld and Mooijaart, 2003). The basic assumption of this model is that the offending pattern of a subject within a certain age strip depends only on a discrete latent variable representing his/her tendency to commit crimes, which follows a first-order homogeneous Markov process. In its current form, however, the model may be applied only in the univariate case, i.e. when the offending pattern of a subject is represented through a single discrete variable. This may be rather restrictive when several offence categories are considered and we wish to take into account that a subject may commit crimes belonging to different categories within the same age strip.

In this paper we show how a latent Markov approach may be also followed to analyse criminal trajectories when offending patterns are represented through a set of binary variables, one for any offence category. As in the latent class model (Lazarsfeld and Henry, 1968), frequently applied to classify subjects according to their criminal behaviour (McCutcheon and Thomas, 1995; Francis *et al.* 2004), we assume local independence, i.e. for any age

strip the response variables are conditional independent given the latent variable. The resulting model will be illustrated in the following Section where we also show how, by restricting appropriately the transition matrix of the Markov chain, it is possible to express hypotheses of interest on the criminal behaviour. Maximum likelihood estimation of this model is dealt with in Section 3 where it is also briefly outlined how we can use the likelihood ratio to choose the number of states of the Markov chain and test hypotheses expressed through restrictions on the transition matrix.

To illustrate our approach we will analyse the criminal trajectories of a cohort of 11,402 offenders born in England and Wales in 1953. Offences are combined into 10 major categories, while criminal careers are aggregated into fixed five-year age periods of the offender's criminal history. The data, drawn from the England and Wales Offenders Index, are publicly available.

2   Multivariate Latent Markov Model

Let X_{tj}, $t = 1, \ldots, T$, $j = 1, \ldots, J$, be a binary variable equal to 1 if a subject is convicted for offence of category j within age strip t and to 0 otherwise; let also $\boldsymbol{X}_t$ be the column vector with elements X_{tj}, $j = 1, \ldots, J$. We assume that, for $t = 1, \ldots, T$, there exists a discrete latent variable C_t such that, given this variable, the elements of $\boldsymbol{X}_t$ are conditional independent. This implies that

$$\phi(\boldsymbol{x}|t) = p(\boldsymbol{X}_t = \boldsymbol{x}|C_t = c) = \prod_{j=1}^{J} \lambda_{cj}^{x_j}(1 - \lambda_{cj})^{1-x_j},$$

where $\lambda_{cj} = p(X_{tj} = 1|C_t = c)$ that, by assumption, is independent of t. We also assume that C_t follows a first-order homogenous Markov chain with transition probability matrix $\boldsymbol{\Pi}$, whose elements are $\pi_{c_1 c_2} = p(C_t = c_2|C_{t-1} = c_1)$, and initial probabilities $\pi_c = P(C_1 = c)$ collected in the vector $\boldsymbol{\pi}$ and that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ are conditional independent given $C_1, \ldots, C_T$. So, we have that

$$p(\boldsymbol{X}_1 = \boldsymbol{x}_1, \ldots, \boldsymbol{X}_T = \boldsymbol{x}_T) =$$
$$\sum_{c_1} \phi(\boldsymbol{x}_1|c_1)\pi_{c_1} \sum_{c_2} \phi(\boldsymbol{x}_2|c_2)\pi_{c_1 c_2} \cdots \sum_{c_T} \phi(\boldsymbol{x}_T|c_T)\pi_{c_{T-1} c_T};$$

in the following, this probability will be denoted by $q(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$.

In order to incorporate in the model hypotheses of interest on the criminal behaviour, we can appropriately restrict the transition matrix $\boldsymbol{\Pi}$. For instance, when the states may be ordered according to the tendency to commit crimes, the hypothesis that offenders begin their careers by committing trivial offences and escalate to more serious crimes later in life may be expressed through the constraint that $\boldsymbol{\Pi}$ is upper triangular. Instead,

the hypothesis that the tendency to commit crimes remain the same for all the life may be formulated by letting $\boldsymbol{\Pi}$ equal to a k-dimensional identity matrix. Fitting the multivariate latent Markov model under this constraint is equivalent to fitting a latent class model that ignores the longitudinal structure of the data.

3   Likelihood inference

Let $\boldsymbol{x}_{it}$ be the observed value of the vector $\boldsymbol{X}_t$ for the i-th subject in a cohort of n subjects. The log-likelihood of the model is then

$$l(\boldsymbol{\theta}) = \sum_{i}^{n} \log q(\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}),$$

where $\boldsymbol{\theta}$ is a short-hand notation of all the parameters. For the maximization of $l(\boldsymbol{\theta})$ we can apply the EM algorithm (Dempster et $al.$, 1977). To describe this algorithm it is convenient to introduce the log-likelihood of the complete data, i.e. the log-likelihood that we could compute if we knew the value of latent variables $C_1, \ldots, C_T$ for all the subjects in the cohort. This function may be expressed as

$$l^*(\boldsymbol{\theta}) \quad = \quad \sum_{c} v_{\cdot 1 c} \log \pi_c + \sum_{c_1} \sum_{c_2} u_{c_1 c_2} \log \pi_{c_1 c_2} +$$
$$\sum_{i} \sum_{t} \sum_{c} v_{itc} \sum_{j} \{x_{itj} \log \lambda_{cj} + (1 - x_{itj}) \log(1 - \lambda_{cj})\},$$

where v_{itc} is a dummy variable, referred to the i-th subject, which is equal to 1 if $C_t = c$ and to 0 otherwise, $v_{\cdot tc} = \sum_{i} v_{itc}$ and $u_{c_1 c_2}$ is the number of transitions from the c_1-th to the c_2-th state.
The EM algorithm alternates the following steps until convergence:

E step. It consists in computing the conditional expected value of the complete log-likelihood, $\tilde{l}^*(\boldsymbol{\theta})$, given the observed data and the current value of the parameters. This is equivalent to compute the conditional expected value of the variables v_{itc}'s and $u_{c_1 c_2}$'s. These expected values, denoted in the following by $\tilde{v}_{itc}$ and $\tilde{u}_{c_1 c_2}$, may be obtained through well-known recursions in the hidden Markov models literature (MacDonald and Zucchini, 1997, Sec. 2.2).

M-step It consists in updating the parameter estimates by maximizing $\tilde{l}^*(\boldsymbol{\theta})$. When the model is unconstrained, this may be simply performed as follows:

$$\lambda_{cj} = \sum_{i} \sum_{t} \tilde{v}_{itc} x_{itj} \Big/ \sum_{i} \sum_{t} \tilde{v}_{itc}, \quad c = 1, \ldots, k, \; j = 1, \ldots, J,$$

$$\pi_c = \tilde{v}_{\cdot 1 c} \Big/ \sum_d \tilde{v}_{\cdot 1 d}, \quad c = 1, \ldots, k,$$

$$\pi_{c_1 c_2} = \tilde{u}_{c_1 c_2} \Big/ \sum_d \tilde{u}_{c_1 d}, \quad c_1, c_2 = 1, \ldots, k.$$

Possible restrictions on $\boldsymbol{\Pi}$ affects only the way in which the elements of this matrix are updated.

To choose the number of latent classes we can rely on a simple procedure based on the likelihood ratio between the model with k states and that with $k + 1$ states, $r_k = -2(\hat{l}_k - \hat{l}_{k+1})$, for increasing values of k. According to this procedure, the optimal number of states, $\hat{k}$, is the smallest k such that the p-value for r_k is greater than a certain threshold, say 0.05. To compute a p-value for r_k we can use a parametric bootstrap procedure based on a suitable number of samples generated from the estimated model with k states. Once the number of states has been chosen, the likelihood ratio may be still used to test hypotheses expressed through restrictions on the transition matrix. In this case we have to compare a model with $\hat{k}$ states restricted according to the hypothesis of interest with the unrestricted model with the same number of states.

References

Bijleveld, C. J. H., and Mooijaart, A. (2003). Latent Markov Modelling of Recidivism Data. *Statistica Neerlandica*, **57**, 3, 305-320.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. series B*, **39**, 1-38.

Francis, B., Soothill, K. and Fligelstone, R. (2004). Identifying Patterns and Pathways of Offending Behaviour: A New Approach to Typologies of Crime. *European Journal of Criminology*, **1**, 47-87.

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

McCutcheon, A. L. and Thomas, G. (1995). Patterns of drug use among white institutionalized delinquents in Georgia. Evidence from a latent class analysis. *Journal of Drug Education*, **25**, 61-71.

MacDonald I. and Zucchini W. (1997). *Hidden Markov and Other Models for Discrete-valued Time Series.* London: Chapman & Hall.

Wiggins, L. M. (1973). *Panel Analysis: Latent Probability Models for Attitudes and Behavior Processes.* Amsterdam: Elsevier.

Application of the modified profile likelihood in stratified models

Ruggero Bellio[1] and Nicola Sartori[2]

[1] Dept. of Statistics, University of Udine, Italy. `ruggero.bellio@dss.uniud.it`
[2] Dept. of Statistics, University Ca' Foscari of Venice, Italy. `sartori@unive.it`

Abstract: In stratified models the modified profile likelihood leads to accurate inference for the parameters of interest, which are common to all strata, eliminating the effect of stratum-specific nuisance parameters. The computation of the modified profile likelihood is simple and leads to substantial improvement over standard likelihood methods, based on the profile likelihood. Here, we propose an application to a negative binomial loglinear model and we compare the results with the case in which the nuisance parameters are modeled as random effects.

1 Introduction

We consider inference in models for independent stratified random variables Y_{ij}, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, such that

$$Y_{ij} \sim p(y_{ij}; \psi, \lambda_i, x_{ij}), \qquad (1)$$

where x_{ij} are explanatory variables. We assume that ψ is the parameter of interest, while $\lambda = (\lambda_1, \ldots, \lambda_k)$ is considered as a nuisance parameter.

In parametric models with parameter $\theta = (\psi, \lambda)$, standard likelihood inference for the parameter ψ is typically based on the profile likelihood, which is the likelihood with the nuisance parameter replaced by its constrained maximum likelihood estimate for fixed ψ. It is well known since Neyman and Scott (1948) that the profile likelihood may lead to very inaccurate inference in stratified models. In particular, this is likely to happen when the number of strata k, which is also the dimension of the nuisance parameter, is large relative to the size of the strata.

In some cases, the solution to this problem is given by means of some inferential separation in the likelihood, as with the conditional likelihood. The conditional likelihood removes the stratum-specific parameters $\lambda_1, \ldots, \lambda_k$, by conditioning on suitable sufficient statistics. As a result, the maximum likelihood estimator and the likelihood-based statistics based on the conditional likelihood have the usual asymptotic properties, as opposed to those

based on the profile likelihood (Andersen, 1970). The problem is that the existence of a conditional likelihood is not guaranteed in a generic model. Here the aim is to propose the use of modified profile likelihood (Barndorff-Nielsen, 1983) as an extension of the conditional likelihood approach in stratified models. There are two major motivations for this. First, when a conditional likelihood is available, the modified profile likelihood is an accurate approximation for it. Second, the modified profile likelihood is a general tool for inference, as the profile likelihood. The theoretical justification for the use of the modified profile likelihood, in place of the profile likelihood, in the presence of many stratum nuisance parameters is given in Sartori (2003). The main point is that, when the number of strata is large compared to the strata sample sizes, the modified profile likelihood has better asymptotic properties than the profile.

Bellio and Sartori (2003) applied the modified profile likelihood in generalized linear models for binary data. Here, after a brief review in Section 2, we consider an application to negative binomial data. A comparison with the random effects model is also considered.

2   The modified profile likelihood

Consider a parametric statistical model with parameter $\theta = (\psi, \lambda)$ and with loglikelihood $\ell(\psi, \lambda)$ satisfying some regularity conditions (Severini, 2000, Chapter 3). The profile loglikelihood is $\ell_{\mathrm{P}}(\psi) = \ell(\psi, \hat{\lambda}_\psi)$, where $\hat{\lambda}_\psi$ is the maximum likelihood estimate of λ when ψ is treated as fixed.

The modified profile loglikelihood (Barndorff-Nielsen, 1983) has the form

$$\ell_{\mathrm{M}}(\psi) = \ell_{\mathrm{P}}(\psi) + M(\psi),\tag{2}$$

where the function $M(\psi)$ is such that $\ell_{\mathrm{M}}(\psi)$ approximates both conditional and marginal loglikelihoods, when they either exist (Barndorff-Nielsen and Cox, 1994, Section 8.2). Remarkably, the modified profile likelihood is quite effective even when neither a conditional nor a marginal likelihood exists. Its main drawback is that the modification $M(\psi)$ is very difficult to compute outside linear exponential families or transformation models. However, recent results in the field of likelihood asymptotics have widened its applicability, and various approximations are now available (Severini, 2000, Chapter 9). In the case of generalized linear models, the version proposed by Severini (1998) is particularly convenient and has modification of the form

$$M(\psi) = \frac{1}{2} \log |j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)| - \log |I_{\lambda\lambda}(\hat{\psi}, \hat{\lambda}; \psi, \hat{\lambda}_\psi)|,\tag{3}$$

where $(\hat{\psi}, \hat{\lambda})$ is the maximum likelihood estimate of the parameters, $j_{\lambda\lambda}$ is the $\lambda\lambda$-block of the observed information, and $I_{\lambda\lambda}$ is given by

$$I_{\lambda\lambda}(\hat{\psi}, \hat{\lambda}; \psi, \hat{\lambda}_\psi) = \mathrm{cov}_{\psi_0, \lambda_0} \left\{ \ell_\lambda(\psi_0, \lambda_0), \ell_\lambda(\psi_1, \lambda_1) \right\} \big|_{(\psi_0 = \hat{\psi}, \lambda_0 = \hat{\lambda}, \psi_1 = \psi, \lambda_1 = \hat{\lambda}_\psi)},$$

$$(4)$$

where $\ell_\lambda(\psi, \lambda) = \partial \ell(\psi, \lambda)/\partial \lambda$ denotes the λ-part of the score function. In the standard asymptotic setting, where the dimension of λ is fixed, likelihood-based inferences based on profile, modified profile and conditional likelihoods are valid to first-order, with no formal improvement for conditional or modified profile likelihoods. Hence, although modified profile likelihood empirically lead to more accurate results, there seems to be no need for such an improvement over the standard method, unless the dimension of the nuisance parameter is large compared to the sample size. A notable instance when this may happen is represented by stratified models, which are considered in the following.

In model (1), the loglikelihood can be written as

$$\ell(\psi, \lambda) = \sum_{i=1}^{k} \ell_i(\psi, \lambda_i), \tag{5}$$

where $\ell_i(\psi, \lambda_i) = \sum_{j=1}^{n_i} \log p(y_{ij}; \psi, \lambda_i, x_{ij})$ is the contribution to the loglikelihood of the i-th stratum.

We note that the presence of stratum-specific nuisance parameters and the independence among strata imply the additivity of the profile loglikelihood. For the same reasons, both $j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)$ and $I_{\lambda\lambda}(\hat{\psi}, \hat{\lambda}; \psi, \hat{\lambda}_\psi)$ are block-diagonal matrices. Hence, also $\ell_\mathrm{M}(\psi)$ is additive, because (3) may be written in the form $M(\psi) = \sum_{i=1}^{k} M_i(\psi)$, where

$$M_i(\psi) = \frac{1}{2} \log |j_{\lambda_i \lambda_i}(\psi, \hat{\lambda}_{i\psi})| - \log |I_{\lambda_i \lambda_i}(\hat{\psi}, \hat{\lambda}_i; \psi, \hat{\lambda}_{i\psi})|. \tag{6}$$

The sample size is $\sum_{i=1}^{k} n_i$ and the dimension of the nuisance parameter is k. In what follows, we assume that the strata are asymptotically balanced, in the sense that each n_i may be written as $n_i = K_i n$, with $A \leq K_i \leq B$ and where A and B are positive finite numbers. When k grows, both sample size and the dimension of the nuisance parameter grow. This is the typical case in which the profile likelihood may fail and the use of conditional or modified profile likelihoods can greatly improve inference. Sartori (2003) studies a two-index asymptotic setting in which both k and n increase to infinity and shows that modified profile likelihood has better asymptotic properties than the profile. In particular, the bias of $\hat{\psi}$ is of order $O(n^{-1})$, while the bias of $\hat{\psi}_\mathrm{M}$, the estimator obtained from $\ell_\mathrm{M}(\psi)$, is of order $O(n^{-2})$. However, results about bias do not give the full picture because they do not take into account the order of standard errors, which depend also on k. On the contrary, sufficient conditions for the usual χ^2 asymptotic distribution of Wald, score and likelihood ratio statistics involve both k and n. The condition is $k = o(n)$ for the profile likelihood, while is $k = o(n^3)$ for

the modified profile likelihood. Hence, unless the strata sample sizes are larger than the number of strata, which is an uncommon practical situation, we cannot expect standard likelihood methods to be reliable. Instead, the modified profile likelihood guarantees accurate inference even in cases with k much larger than n.

3   Negative binomial loglinear model for count data

The Poisson loglinear model is a classical model for count data, but often overdispersion is present. A common choice to handle it is to resort to the negative binomial model; a gentle introduction is given in Venables and Ripley (2002, §7.4). It is well known that there is not a unique way for specifying the negative binomial loglinear model (see Lindsey, 1999). Here, we assume that the marginal distribution of the response Y_{ij} has mean and variance

$$E(Y_{ij}) = \mu_{ij} = \exp(\lambda_i + x_{ij}^T \beta)\,, \qquad\qquad V(Y_{ij}) = \mu_{ij} + \frac{\mu_{ij}^2}{\alpha}\,. \quad (7)$$

The parameter α determines the amount of overdispersion, while the intercepts λ_i deal with the stratified structure.

As an example of application, we consider the *Epileptic seizures* data of Thall and Vail (1990), which are also included in the R library MASS (Venables and Ripley, 2002). The data come from a longitudinal study on epileptics. A group of 59 patients were observed for a baseline period of 8 weeks and then randomized to a treatment for four successive two-week treatment periods; the response was the number of observed seizures. Venables and Ripley (2002, §10.4) report two possible ways of analysing the dataset, and in both cases the Poisson fit indicates the presence of substantial overdispersion. Here we focus on the case which uses a loglinear model with several predictors, including log-baseline counts, treatment status and the indicator of the fourth visit (V4). The total sample size is given by 59×4 observations. Note that all predictors but V4 are time invariant, thus they are confounded with subjects and their effects can not be estimated in models with subject-specific fixed intercepts. However, the modified profile likelihood allows to study the evolution of the response over time and the amount of overdispersion, removing any unobservable individual heterogeneity. For the sake of comparison, we also present the maximum likelihood estimates obtained from a random intercepts model, assuming a Gaussian distribution for λ_i. Table 1 reports the results.

We note very similar estimates of the coefficient of V4 with all methods but, more importantly, a quite different indication about the degree of overdispersion from the profile likelihood and the modified profile likelihood. It is somehow reassuring that the estimate of α from the Gaussian random effects model is close to that from the modified profile likelihood.

TABLE 1. Epileptic seizures, parameter estimates with different methods.

Method	Estimates (s.e.)	
	V4	Index (α)
Profile Likelihood	-0.12 (0.08)	13.84 (3.53)
Modified Profile Likelihood	-0.11 (0.09)	7.46 (0.94)
Gaussian Random Effects	-0.12 (0.09)	7.40 (0.95)

References

Andersen, E.B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society, Series B*, **32**, 283-301.

Barndorff-Nielsen, O.E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **70**, 343-365.

Barndorff-Nielsen, O.E. and Cox, D.R. (1994). *Inference and Asymptotics*. London: Chapman & Hall.

Bellio, R. and Sartori, N. (2003). Extending conditional likelihood in models for stratified binary data. *Statistical Methods & Applications*, **12**, 121-132.

Lindsey, J.K. (1999). On the use of corrections for overdispersion. *Applied Statistics*, **48**, 553-561.

Neyman, J. and Scott, E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, **16**, 1-32.

Sartori, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika*, **90**, 533-549.

Severini, T.A. (1998). An approximation to the modified profile likelihood function. *Biometrika*, **85**, 403-411.

Severini, T.A. (2000). *Likelihood Methods in Statistics*. Oxford: Oxford University Press.

Thall, P.F. and Vail, S.C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics*, **46**, 657-671.

Venables, W.N. and Ripley, B.D. (2002). *Modern Applied Statistics with S (Fourth Edition)*. New-York: Springer-Verlag.

Analysis of Breast Cancer Survival Data with missing information on stage of disease and cause of death

Bellocco Rino[1]

[1] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet,
POBOX 281, Stockholm, Sweden

Abstract: Aim of this paper is to study whether social class is related to breast cancer survival, in a cohort of 4709 breast cancer patients diagnosed in Sweden in 1993 and followed until the end of 2001, while adjusting for possible demographics and tumor related confounders. The data are provided by the Swedish Cancer Registry and are matched to the death registry by using the unique Swedish Personal Registration Number.

The statistical problem is that the most recent cases have not reported in the registry, as far as it concerns with the underlying cause of death, and standard cause specific survival analysis will turn to exclude those patients, then affecting our ability to detect any statistical difference in the effect of our covariate of interest. Furthermore, a related problem is that for some cases some important covariates (tumor stage) are missing, due the fact that the regional cancer registries have not provided the requested information.

In this application simple missing data imputations have been incorporated into a standard survival data analysis problem, based on the estimation of the Kaplan-Meier estimator and Cox proportional hazards regression model.

As the type of failure is truncated by time, imputing the cause of death will increase the follow-up time, therefore allowing to best study the survival distribution. Moreover, when also a confounder is missing completely at random, it is possible to detect the effect of the main exposure variable with more accuracy.

Keywords: Survival Analysis; Missing Data ; Imputation; Social Class

1 Introduction

Epidemiological findings indicate that breast cancer survival is related to socioeconomic factors. Women of lower socioeconomic status have generally been found to have poorer survival.

Epidemiological findings indicate that both breast cancer incidence and survival are related to socioeconomic factors. Women of lower socioeconomic status are at lower risk of developing breast cancer (Faggiano et al.) but tend to have poorer survival compared to socioeconomically more favored women (Vågerö & Persson).

A common problem in analysis of survival data is the presence of competing risk. When the cause of death is known, it is possible to study the effect of covariates on cause-specific hazards by treating the deaths from other causes as censored observations in a Cox regression model (Cox & Oakes). As the follow-up increase, the time available for quality checking of the death certificates decreases and therefore the statistician has to face the dilemma whether to censor the data at an earlier period of time, where complete information on the endpoint is fully available, or to try using all the data by imputing the missing value of cause of death (Andersen et al). Furthermore, even if complete information on social-economic status is present, it is possible that for the same reason some possible covariate, such as tumor stage, might be missing for a particular reporting center.
Therefore, we propose a simple strategy to incorporate the two components of missing data in the analysis, under the simplifying assumption that missingness is completely at random, in the standard survival analysis procedures.

2   Material and Methods

This underlying study is based on a linkage between the following Swedish population-based registers: the Cancer Register, five Regional Cancer Registers, the 1970, 1980, 1985 and 1990 Census databases, the Fertility Register, Emigration Register, and Cause of Death Register. Record linkages were made possible by using the individually unique National Registration Number (NRN) assigned to each resident in Sweden at the time of birth or residency. These are high quality registries: In 1993, 99% of the breast cancer cases were morphologically or cytologically verified and the overall reporting to the Cancer Register was estimated to be about 98% of all diagnosed cases (National Board of Health and Welfare). A validation study of breast cancer reporting from one Swedish hospital showed that only 1% of all diagnosed cases were missing in the register during the period 1971-1991 A total of 4645 women were diagnosed with invasive breast cancer as first diagnosis from January 1 to December 31 in Sweden in 1993. Of these, 1646 (35%) women have died as of December 31, 2001, the end of the follow-up period. However, 298 women died after December 31, 1998, the date after which the cause of death was unknown. The total number of women with ascertained cause of death was 1348, and 772 of these deaths (57.3%) were due to breast cancer.
Standard survival analyzes are performed: the survival distribution is estimated by Kaplan-Meier technique, and log-rank test is used to assess the influence of the main exposure variable. We also run proportional hazard regression model to study how the estimates change according the different scenario of missing data for the covariates.
Imputation of missing cause of death was done in two steps: first we a logistic regression model, in which for a woman with known cause of death

FIGURE 1. Partial Follow-up.

we model the logit of the probability of dying of breast cancer, given the covariate patterns (marital status, age, region of diagnosis). The second step, for a woman with missing cause of death is to generate a binary random variable with mean given my the fitted probability.

3   Results

In figure Figure 1 we show the failure distributions when we end the follow on the first date (December 31, 1998); the log-rank test shows that the two survival distributions are statistically different with a p-value = 0.01. We also observe that more than 80% of women diagnosed with cancer are still alive after 6 years of follow-up.

In figure Figure 2 I show the same distribution after multiple imputation of cause of deaths has been performed and median values of the estimated failure distributions have been calculated. Not surprisingly the log-rank test shows an even higher statistical difference, (P-value =0.002). It is also important to notice that apparently the hazard of dying of breast cancer for high social class women seems to level off after 8 years from diagnosis, whereas the hazard for low social class women seems being constant.

In the second stage of our missing data problem, I considered the effect of
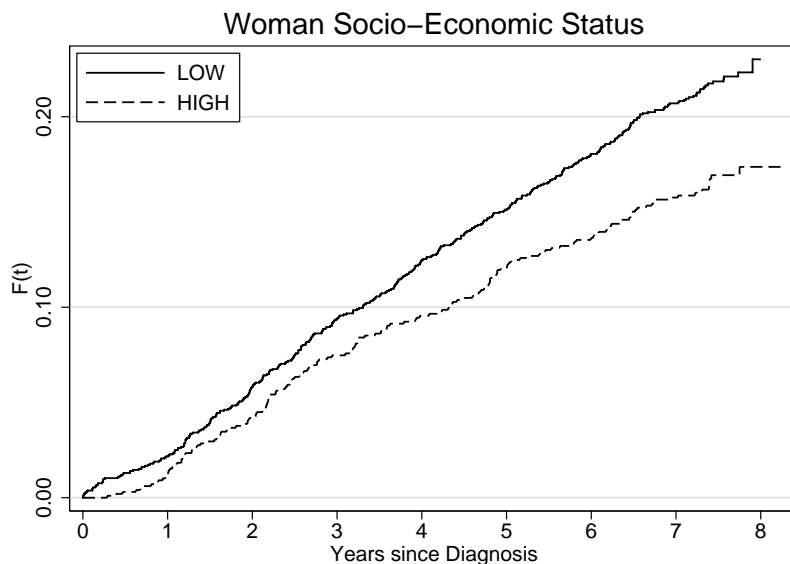
FIGURE 2. Complete Follow-up.

TABLE 1. Social-Economic Effect adjusted by tumor stage: Hazard Ratio, 95% Confidence Intervals (CI), P-values .

	Model 1: Stage Available Data	Model2: Stage Imputed Data
hazard ratio	0.81	0.75
95% CI	0.65- 1.01	0.62-0.90
P-value	0.06	0.02

tumor stage, as a possible confounder for the relationship between social status and time to death of breast cancer. Tumor stage was missing for one of the regional cancer registries in Sweden and as many as 1200 women would not be considered in the final model.

In Table 2 we report the results from fitting two different models: model (1) is considering only patients with available tumor data, model (2) is taking into account the missing component of the covariate, according to the the simple missing data indicator method (Greenland and Finkle).

4   Conclusions

Preliminary results show that it is possible to incorporate missing data into a standard survival data analysis. Multiple imputation of the failure indicator might increase the ability of detecting significant differences between survival distributions, as we increase the follow-up time. I have also compared the observed results with the Kaplan-Meier estimator when considering any type of death as the the endpoint of the study and some conclusions can be drawn. As far it concerns with the imputation of the tumor stage, although the method might produce some severe biased results in some cases, in this situation it is reasonable to assume it might affect our results, as both missing data can be easily completely at random and and only affecting the confounder of interest.

References

Andersen, J., Goetghebeur, E., Ryan, L.. (1996). *Missing Cause of death information in the analysis of survival Data.* **15**, 2191-2201.

Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data.* Chapman and Hall: London.

Faggiano, F., Partanen, T., Kogevinas, M., Boffetta, P. (1997). Socioeconomic differences in cancer incidence and mortality. *IARC Scientific Publications*, **138**, 65-176.

Garne, J.P., Aspegren, K., Moller, T. (1995). *Validity of breast cancer registration from one hospital into the Swedish National Cancer Registry 1971-1991.* Acta Oncologica, 34(2):153-6.

Geenland, S., Finke, W.D. (1995). *A critical look at methods for handling missing covariates in epidemiologic regression analysis*, **142(12)**, 1255-1264.

Vågerö , D., Persson, G. (1987). Cancer survival and social class in Sweden. *Journal of Epidemiology and Community Health.*, **41(3)**, 204-9.

National Board of Health and Welfare (1996). Cancer incidence in Sweden 1993. *Centre for Epidemiology, National Board of Health and Welfare.* Stockholm, Sweden.

A split-plot analysis for microarray experiments

Rossella Berni and Federico M. Stefanini

[1] Department of Statistics "G. Parenti" University of Florence

Abstract: We focus on a split-plot analysis for microarray experiments to account for the rich hierarchical structure typical of this measurement process. The real operative levels in the experimentation are here addressed. In particular, the levels of gene factor are reduced performing a selection based on variability of the intensity. Further issues here considered are the distinction between random and fixed effects and the consideration of the diameter as spot's covariate.

Keywords: Split-plot design, microarray experiments, spot effect, robust design.

1 Introdution: split-plot designs and microarrays

The aim of this work is to investigate the applicability of split-plot designs in a simple experimental setup. More precisely, it is well known in literature the role of the split-plot design as a plan for robust product experimentation (Box and Jones, 1992). In fact, the specific structure (framework) of a split-plot can be easily arranged in order to take care of the external variability and, also, of the hierarchy among factors according to operative levels, in particular, whole and sub-plot. External variability is a concept connected to the definition of environmental variables or, also, noise factors, even though measurable and controllable. In microarray experiments, the concept of external variability can be assigned to the array and print-tip (pin) factors. Therefore, the set of factors of interest, also called internal factors of the process, are the variables directly influencing the intensity measure and the gene expression.

Operative levels are fundamental characteristic of a split-plot design (Logothetis and Wynn, 1990). With microarrays, we suppose three operative levels: a "slide" level, a primary level in which we consider the array factor, the pin factor and the correspondent interaction; a secondary level, with the factor of interest as gene, dye, variety and the related crossproducts; a third level, which we could call "spot" level, by which we attempt to measure the effect due to the physical features of spots.

Regarding these issues, we must consider the following problems. First of all, we build a split-plot design for data just collected, so we perform a split-plot analysis, only considering the related model applied to our data.

Secondly, this split-plot analysis must take care of crossproducts between factors belonging to different operative levels. For example, the interaction between the array and the gene factors. In this work, this aspect must be evaluated also considering the nature of the variables involved. At each operative level, experimental factors could be random or fixed factors.

In general, in microarray experiments, array pin and gene are considered as random factors. In our application, we consider array as fixed factor, pin as random factor, the spot covariates as random factors. Furthermore, the gene factor is evaluated as a fixed factor at an initial step of the analysis but, in order to reduce the number of levels (type of genes), we make a selection of genes based on a measure of variability for the fluorescence intensity. Consequently, by the use of this transformed gene factor, we suppose that genes are similar, or homogeneous, as regards the fluorescence variability. This assumption has to be weakened in future work.

Another feature is about the spot covariates. In general, it is well known the difficulty to evaluate the "spot" effect, just because the measures related to the spot are affected by the background noise. Consequently, auxiliary spot's indices, such as uniformity, circularity and diameter, are crude estimates. Nevertheless we apply a spot analysis by considering two possible approaches: the average of each spot variable calculated within the pin factor, here confounded with the sub-array factor; otherwise by considering the three replicated spots for the same gene.

2   The suggested model

The model here proposed could be considered a general model for split-plot analysis in the microarray field. Here two arrays were considered, arranged in a dye-swap scheme. The layout of the experiment is made by two target samples of maize ear tissues: a wild type genotype and a mutant genotype. There are 8 grids (subarrays) in a 4 by 2 lattice, and each grid is a square of 45 by 45 spots. Detailed explanations about the array manufacturing can be found at the URL address http://www.zmdb.iastate.edu/ on internet, array batch number 605.03.

The model has the following general expression:

$$y_{ijkl} = \mu + r_l + E_j + \eta_{jl} + D_i + (DE)_{ij} + \psi_{ijl} + S_k + (ES)_{jk} + (DS)_{ik} + e_{ijkl} \quad (1)$$

where, for simplicity, the letters E, D, S stay for the three operative levels of the split-plot; Environmental, Design and Spot level. For each of these levels we have a set of variables; y_{ijkl} is the response for the lth replicate of the ith level of factor D, the jth level of factor E and the kth level of factor S; the r_l term is the random effect of the lth replicate with $r_l \sim N(0, \sigma_r^2)$. In the E set we consider: array, pin and the interaction $array * pin$; in the D set we put: gene, colour, channel, and the interactions $gene * channel$, $array * colour$, $gene * colour$; in the set S are considered the variables of the spot:

circularity, uniformity and diameter, and, eventually, their interaction with the array factor. We don't consider the crossproducts among the gene factor and the spot measures. We must point out that the terms of the model (1): η_{jl}, ψ_{ijl}, and e_{ijkl}, represent the independent error components, supposed Normally distributed with null expected value and proper variance. In the next section a first empirical example is applied; the two proposed models are simpler than (1): regarding the level of the split-plot design and the number of factors involved; in addition, we consider three replicates for the same gene and we evaluate only one spot covariate: the diameter.

Our approach builds on usual anova models (Churchill et al., 2000) but it is devoted to an improved exploitation of information about the measurement process, both in external and internal noise factors. The suggested class of models differs from Wolfingers' (Wolfinger et al., 2001) two-step procedure in which one-at-a-time gene analysis is performed.

As regards the case study, two models are proposed following a two-levels split-plot design in which the *array* factor, considered as a fixed factor, is arranged as whole-plot variable; the print-tip factor (here called *PIN* factor) is considered as a whole-plot classification factor at random effects nested within the *array* factor, while *gene* and *channel* are assigned to subplots. The *gene* factor is considered as a fixed factor,the *channel*, (here confounded with colour), is a fixed factor. The levels of the gene factor are reduced by genes selection: the procedure selects 96 genes which show large fluorescence differences between dyes (6351 observations).

Furthermore, two error components are defined: the first is related to the *array* and *PIN* factors, while the second is a pooled error formed by the residual terms of higher order of the subplots and the interactions between the terms of the subplots and the classification effects.

Considering the formula(1) in section 2, for the first model, we put in the E-set the *array* and *PIN* factors, in the D-set we insert the *gene* and *channel* factors and two interactions: $array * channel$ and $gene * channel$. The second model the *diameter* as spot covariate. This variable is considered as a continuos factor at random effects nested within the *array* factor.

For the first model, tables (1) and (2) show the results for random and fixed effects. The convergence criteria are met at the second iteration, using REML as estimation method.

The fixed effects are significant, but the interaction $gene * channel$; (table (2)). The tests are computed using the Type III SS, to take into account of the unbalanced design.

The second model including the *diameter* of the spot is also satisfactory. Regarding *diameter* as a continous factor at random effects. The convergence criteria are met at ninth iteration and *diameter* is highly significant within each array. Tables (3)and (4) show the results for the random and fixed effects.

The results for the fixed effects (table (4)) are similar to the results obtained

TABLE 1. Solution for Random Effects - I model

Effect	block	PIN	Est	std Err	t-test	p-value
PIN	1	1	0.1174	0.09187	1.28	0.2012
PIN	1	2	-0.2805	0.09169	-3.06	0.0022
PIN	1	3	0.1803	0.09183	1.96	0.0496
PIN	1	4	-0.01718	0.09161	-0.19	0.8512
PIN	2	1	-0.06318	0.09189	-0.69	0.4918
PIN	2	2	-0.09721	0.09179	-1.06	0.2896
PIN	2	3	-0.04036	0.09188	-0.44	0.6605
PIN	2	4	0.2007	0.09159	2.19	0.0284

TABLE 2. Results for fixed effects of interest- test F (df) and p-values - I model

Effect	df	Mean Square	F-value	p-value
Array	1	19.89	0.88	0.4169
I error	3	22.54	-	-
Channel	1	9.72	14.88	0.0001
Array*Channel	1	58.14	89.04	$< .0001$
gene	95	238.36	365.04	$< .0001$
gene*Channel	95	0.08	0.12	*n.s.*
II error	6151	0.65297	-	-

TABLE 3. Solution for Random Effects - II model

Effect	block	PIN	Est	std Err	t-test	p-value
PIN	1	1	0.1186	0.06551	1.81	0.0702
PIN	1	2	-0.1760	0.06538	-2.69	0.0071
PIN	1	3	0.1289	0.06548	1.97	0.0491
PIN	1	4	-0.07155	0.06525	-1.10	0.2729
Diameter	1	-	-0.05375	0.00192	-28.01	$< .0001$
PIN	2	1	-0.01518	0.06555	-0.23	0.8169
PIN	2	2	-0.03625	0.06544	-0.55	0.5797
PIN	2	3	-0.06145	0.06522	-0.94	0.3483
PIN	2	4	0.1129	0.06526	1.73	0.0837
Diameter	2	-	-0.05838	0.00186	-31.46	$< .0001$

by the first model.

TABLE 4. Results for fixed effects of interest- test F (df) and p-values - II model

Effect	df	Mean Square	F-value	p-value
Array	1	3.18	0.20	0.6843
I error	3	15.823	-	-
Channel	1	13.04	23.57	< .0001
Array*Channel	2	82.60	149.26	< .0001
gene	95	39.27	70.96	< .0001
gene*Channel	95	0.14	0.26	n.s.
II error	6150	0.5534	-	-

3   Concluding remarks

It is relevant to note that this is a first attempt to analyze this kind of data using a split-plot model.Therefore,these are preliminary results to be revised towards the consideration of a third level of the split-plot. In fact, given the relevance of the assignment of factors to the level of the split-plot design, we point out that this aspect must be notably improved.
The possibility of heterogeneous variances among genes should also be addressed as the key issue in further work.

References

Box G.E.P., Jones S. (1992). Split-plot design for robust product experimentation. *Journal of Applied Statistics,* **19(1).**

Kerr M.K., Martin M. and Churchill G.A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology.* **7,** 819-837.

Logothetis N., Wynn H.P. (1990). *Quality Through Design.* Oxford: Clarendon Press.

Wolfinger R.D., Gibson G., Wolfinger E.D., Bennett L., Hamadeh H., Bushel P., Afoshari C., Paules R. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology.* **8(6)** 625-637.

Comparison between the Parametric Mixing Distribution with Mover-stayer Model and the Nonparametric Mixing Distribution for the Analysis of Tower of London data

Md Azman Shahadan and Damon Berridge

[1] Centre For Applied Statistics, Fylde College, Lancaster University, LA1 4YF, Lancaster, England, United Kingdom

Abstract: The Tower of London data have one serious problem. The problem is the high proportion of stayers, because parametric estimation of random effects tends to underestimate the number who are stayers. In this paper, we will use alternative estimation procedures like the parametric mixing distribution with mover-stayer model and the non-parametric mixing distribution methods. In this paper we try to answer how well these alternative procedures compare with each other?

Keywords: Random effects; Endpoints; NPML; Mover-stayer; Tower of London.

1   Introduction and Motivation

Being able to plan efficiently is important in many of the complex behaviours of life such as organising work schedules, making travel plans or even preparing meals (Shallice, 1982). In order to study shortcomings in executive planning, Shallice (1982) developed the Tower of London (TOL) task. Since the publication of Shallice's research, the TOL task has been used extensively as a test of planning ability in both adult and young child populations. Despite the value of the TOL in the assessment of executive planning, a review of the existing tower systems suggested that several changes are needed to adapt them for use with young children. In this paper, we use the datasets and the TOL tasks format are provided by Shimmon & Lewis (2003). These experimental datasets are concerned with binary repeated measures on the TOL task applied to young children (testing their planning ability to solve problems) at three different times. The TOL experiment consists of a series of tasks which, in this study, were carried out repeatedly over time. In this series of tasks, the subjects provide us with a sequence of binary responses, where 1 indicates success and 0 means failure, in any particular task.
At time 1, 115 children were recruited from pre-school playgroups in rural areas around Lancaster. The same children were then tested six months

later in a second wave of data collection. The final wave of testing took place 12 months after phase 1. Many of the 30 children who discontinued participation at time 2 and time 3 did so due either to the failure of their parents to return permission slips allowing further participation or to the departure of those particular children from the area of study (Lancaster). Since the missing cases or the drop-outs are ignorable, we exclude the data for all the missing cases from our analyses and analyse only the complete sequences in the data set.

In reviewing the literature, we have found that, besides yielding the sequences of binary response variables, the experiments also involve a number of factors. Children's ability to inhibit salient aspects of the environment has some effect. The Stroop day/night task is designed specifically for young children by way of testing their ability to inhibit salient aspects of the environment. Children must inhibit their natural inclination to respond "day" when presented with the sun and "night" to the moon by saying the opposite of what they see (night to the sun and day to the moon). Children who get high scores on the Stroop day/night tasks tend also to success in the TOL tasks. Language ability also plays an important role in performing the TOL tasks. The British Picture Vocabulary (BPV) test was used to test children's verbal language ability in monthly units. In order to achieve success with TOL tasks, children have to understand the verbal language command or instructions. The last kind of factor or explanatory variable is the child's early stage of mind development. The false-belief task was used to discover how the theory of mind works. In these experiments, four kinds of false-belief task are used. All four false-belief tasks are the unexpected contents task, the unexpected transfer with the Sally-Anne task, the visual ambiguity task and the appearance reality task. Each false-belief task scores 1 or 0. Scores are added together to get the final score for false-belief. The main motive for analyzing these data is the presence of stayers. The mover-stayer model assumes that each subject is either a "mover" or a "stayer", and that stayers do not move (zero or very low probability of change).

2   Random Effects Models

The random effects model is a particular example of a mixed model that is widely used for the analysis of longitudinal data. Let subject i be observed at time t, then the effects of covariate x_{it} on the outcome y_{it} can be represented in the logistic regression model; $log\{\frac{P(y_{it}=1)}{1-P(y_{it}=1)}\} = \beta'x_{it} + \varepsilon_i$, where $y_{it} = 1$ for a successful outcome, 0 otherwise. The random effects are assumed to be normally distributed with zero mean and variance σ_ε^2. The above random effects model can be estimated using parametric estimation. See, for example, Fahrmeir & Tutz (2001).

3   Parametric Mixing Distribution (PMD) with Mover-Stayer Model (MSM)

The mover-stayer model (MSM) can be incorporated into the parametric estimation of a random effects model. A degree of flexibility (to include the MSM) can be achieved if we represent the proportions of stayers as end points; the likelihood will then be obtained as follows:

$$L_i(\beta, \sigma, \psi_0, \psi_1) = \frac{\psi_0}{1 + \psi_0 + \psi_1}[\prod_{t=1}^{T_i}(1-y_{it})] + \frac{\psi_1}{1 + \psi_0 + \psi_1}[\prod_{t=1}^{T_i} y_{it}] + \frac{L_i(\beta, \sigma)}{1 + \psi_0 + \psi_1}$$

(1)

where ψ_0 and ψ_1 are unknown but can be estimated at the end-points as parameters, and L_i is the sequence likelihood. The estimated proportion of stayers in state zero is given by $p_0 = \frac{\psi_0}{1+\psi_0+\psi_1}$ and the estimated proportion of stayers in state one is given by $p_1 = \frac{\psi_1}{1+\psi_0+\psi_1}$. For a more detailed discussion of the MSM, please refer to Barry et al. (1989).

4   Nonparametric Mixing Distribution (NPMD)

The Generalised Linear Latent and Mixed Model (GLLAMM) program in STATA for mixture distributions, written by Rabe-Hesketh et al. (2002), has an option of using nonparametric maximum likelihood (NPML) for parameter estimation. In order to avoid the specification of a parametric form for the mixing distribution, a nonparametric approach, based on a finite mixture, is considered.

The NPML estimate of $G(Z_i; \beta)$, when it exists, is known to be a discrete distribution on a finite number, K, of mass-points, with masses π_k at locations $z_k, k = 1, ..., k$. Thus the profile likelihood in β, maximized over $G(.)$, is the K-component finite mixture log likelihood

$$\ell = \sum_{i=1}^{n} log(\sum_{k=1}^{K} \pi_k f(y_i|z_k, \beta))$$

(2)

where K, z_k and π_k are functions of β. In order to maximize this profile likelihood, we can reformulate the problem as the maximization of the joint likelihood

$$\ell(\beta, K, \pi_1, ..., \pi_{k-1}, z_1, ..., z_k) = \sum_{i=1}^{n} log(\sum_{k=1}^{K} \pi_k f(y_i|z_k, \beta))$$

(3)

over β and all the parameters of the mixture distribution. The number of components K is unknown, so the log likelihood too has to be maximized.

It is clear that because of the mass-points in NPML being freely locatable, it is possible to take into account the parameters for the endpoints at $\pm\infty$. The number of locations (mass-points) is increased until the likelihood is maximised.

5   Results and Conclusions

Table 1: Regression estimates and standard errors (in parentheses) for the PMD, PMD + MSM and NPMD methods respectively used on the TOL data.

Parameter	PMD	PMD + MSM	NPMD
Quadrature points/ Mass-points	20	20	3
-2loglikelihood: Null model	987.145	987.145	987.145
-2loglikelihood: Parsimonious model	661.590	650.894	650.988
AIC	671.590	662.894	662.988
β_0	-6.954	-6.199	-8.211
	(0.815)	(0.835)	(0.898)
β_1(Language Ability)	0.067	0.065	0.072
	(0.014)	(0.013)	(0.014)
β_2 (False Belief Task)	0.366	0.375	0.359
	(0.110)	(0.102)	(0.100)
β_3 (Stroop Day/Night Task)	0.088	0.086	0.072
	(0.033)	(0.030)	(0.027)
Scale Parameter	2.129	1.234	-
(ω)	(0.278)	(0.251)	-
Mass-point 1/	-	0.291	Fixed
End-point 1	-	(0.101)	-
Proportion	-	0.227	0.235
Mass-point 2/	-	Fixed	2.896
End-point 2	-	-	(0.579)
Proportion	-	-	0.397
Variance Component	4.509	1.523	11.964
	(1.181)	(0.249)	-

The Akaike Information Criteria (AIC) indices of parametric mixing distribution with mover-stayer model (PMD + MSM) and nonparametric mixing distribution(NPMD) in table 1 show us that the PMD + MSM and NPMD estimate similar models. In the PMD + MSM, the stayers (either all failures or all successes) are captured by endpoint estimation, whereas in the NPMD method, the stayers are estimated by the method of free finite mixture (mass-points). The results also indicate that the parametric approach underestimates the magnitude of the mover-stayer problem. It is clear that the tail behaviour of the normal distribution is inconsistent with "stayers"(Barry et al., 1989).

The -2loglikelihood and AIC show that the PMD + MSM and the NPMD model are better in terms of estimation compared to the PMD model. The PMD + MSM and NPMD models take into account the high proportion of stayers.

The PMD + MSM and NPMD approaches cope equally well (deviances and AIC). However, the NPMD approach seems more efficient in terms of the number of mass-points required to specify the mixing distribution (3 mass-points of NPMD compared to 20 quadrature points for PMD + MSM). Moreover the NPMD approach is computationally much less intensive than the parametric approach.

References

Barry, J., Francis, B., Davies, R., and Stott, D. (1989) *SABRE: Software for the analysis of binary recurrent events.* Lancaster University: Centre for Applied Statistics.

Fahrmeir, L. and Tutz, G. (2001) *Multivariate statistical modelling based on generalized linear models (2nd edition).* N.Y.: Springer.

Rabe-Hesketh, S., Pickles, A., and Skrondal, A.   (2001) *GLLAMM manual.* Technical report 2001/02. University of London: Department of Biostatistics, Institute of Psychiatry, King's College.

Shallice, T. (1982) Specific impairments of planning. *Philosophical Transactions of the Royal Society of London B, Vol.298*, 199-741.

Shimmon, K. and Lewis, C. (2003) *Studies on executive function.* Unpublished research. University of Lancaster: Department of Psychology.

PH and Non-PH Frailty Models for Multivariate Survival Data

M. Blagojevic[1] and G. MacKenzie [1]

[1] Keele, University, Centre for Medical Statistics, Keele, Staffordshire ST5 5BG, UK

Abstract: We generalize the previously developed Non-PH CTDL-Gamma and the PH Weibull-Gamma frailty models to correlated survival data. In particular, we seek analytical results using the marginal approach, to determine whether the univariate results generalize to the multivariate context. We consider both the shared and correlated frailty cases. We also develop non-parametric frailty models which enable us to check the appropriateness of the assumed distributional form of the random effect.

Keywords: Frailty Models; Non-PH; EM; Finite Mixtures; Correlated Survival Model.

1 Introduction

In our earlier work, we have extended the Weibull proportional hazards (PH) regression survival model to a Gamma frailty model by means of a multiplicative random effect acting on the hazard function (Hougaard 1984). However, not all survival data are PH and it is therefore useful to explore alternative models which are non-PH. This is relevant as, increasingly, random effect models are being used to analyze multivariate survival data (Ha 2001).

A flexible non-PH model is the Canonical Time-Dependent Logistic (CTDL) described by MacKenzie (1996) and later by MacKenzie(1997). We have already generalized this model by including a multiplicative Gamma frailty term in the hazard function. The resulting frailty model was obtained in closed form and we compared its properties with the Weibull frailty model, noting the connection with a general class of frailty models described by Aalen (1988). The performance of the four models, Weibull and CTDL with and without frailty, was investigated using data from the N. Ireland lung cancer study and it was shown that the CTDL-Gamma model provided the best fit. In addition, non-parametric frailty models are developed to check whether Gamma distribution is appropriate for the random effect. We now extend the models to the multivariate case and special consideration is given to the correlated frailty scenario.

2   Univariate Survival Models

A non-PH model, the CTDL regression model (MacKenzie 1996), is defined by the hazard function

$$\lambda(t|x) = \lambda p(t|x), \tag{1}$$

where $\lambda > 0$ is a scalar, $p(t|x) = \exp(t\alpha + x'\beta)/\{1 + \exp(t\alpha + x'\beta)\}$ is a linear logistic function in time, α is a scalar measuring the effect of time, β is a $p \times 1$ vector of regression parameters associated with fixed covariates $x' = (x_1, \ldots, x_p)$ and $\theta' = (\lambda, \alpha, \beta)$.

When developing the CTDL-gamma mixture model, we assumed that the random component has a multiplicative effect on the hazard, such that $\lambda(t; x, u) = u\lambda(t; x)$. U follows a Gamma distribution with $E(U) = 1$ and $V(U) = \sigma^2$. We then used the marginalization approach to obtain the pdf for the resulting marginal frailty distribution:

$$f_f(t|x) = \frac{\lambda p_i}{\left\{1 - \frac{\lambda\sigma^2}{\alpha}\log_e(g_iq_i)\right\}^{1+\frac{1}{\sigma^2}}} \tag{2}$$

where,

$$\begin{aligned} p_i &= \exp(t_i\alpha + x_i'\beta)/\{1 + \exp(t_i\alpha + x_i'\beta)\} \\ q_i &= 1/\{1 + \exp(t_i\alpha + x_i'\beta)\} \\ g_i &= 1 + \exp(x_i'\beta) \end{aligned} \tag{3}$$

and where, for notational convenience, we have suppressed the dependence on time and the covariates on the LHS of (3).

Similarly for Weibull-gamma model, we found that:

$$f_f(t|x) = \frac{\lambda^\rho \rho e^{x'\beta} t^{\rho-1}}{\left\{1 + \sigma^2(\lambda t)^\rho e^{x'\beta}\right\}^{1+\frac{1}{\sigma^2}}} \tag{4}$$

3   Non-Parametric Frailty

The estimated effect of covariates may be influenced (to a varying degree in different sets of data) by the choice of the distributional form of the frailty density. In order to minimize the impact of frailty distribution assumption, we fit a non-parametric (NP) frailty component based on a finite mixture. We use the EM algorithm for implementation. We are interested in estimating the NP frailty component simultaneously with the mixing proportions. These estimated values will typically suggest the mixture from which the data were generated and hence will provide a useful check on any parametric assumptions made.

The resulting CTDL and Weibull log-likelihoods are, respectively:

$$\ell_{ctdl}(\pi,\theta) = \sum_{j=1}^{c}\sum_{i=1}^{n}\left\{ z_{ij}\log_e\pi_j + z_{ij}[\delta_i\log_e(u_jp_i) + \frac{u_j\lambda}{\alpha}\log_e(q_ig_i)]\right\} \quad (5)$$

where, π_j is the jth component of the mixture, c is the dimension of the mixture, $u_j = e^{c^{ij}\gamma}$ is the non-parametric frailty component, θ is the vector of parameters to be estimated, p_i, q_i, g_i are as before, and

$$\ell_w(\pi,\theta) = \sum_{j=1}^{c}\sum_{i=1}^{n}\left\{ z_{ij}\log_e\pi_j + z_{ij}[\delta_i\log_e(\lambda^\rho\rho t_i^{\rho-1}e^{x_i'\beta}u_j) - (\lambda t_i)^\rho e^{x_i'\beta}u_j]\right\} \quad (6)$$

An algorithm was written in S-Plus (V4.5) to maximize (5) and (6).

4 Multivariate Survival Data

We turn now to the idea of generalizing the parametric frailty models introduced earlier to correlated survival data. In particular, we seek analytical results using the marginal approach, in order to determine whether the univariate results generalize to the multivariate context.

Suppose we have $f(t_i|u_i,\theta)$ and $g(u_i|\sigma^2)$ where $t_i = (t_{i1}, t_{i2}, \cdots, t_{i_{m_i}})$ is the vector of survival times on the ith subject. m_i is the number of measurements on the ith subject, whence t_{ij}, $i = 1,\cdots,n$; $j = 1,\cdots,m_i$, become our data. The joint likelihood of t and u is then:

$$L(\theta,\sigma^2) = \prod_{i=1}^{n} f(t_i|u_i,\theta)g(u_i|\sigma^2) \quad (7)$$

However, under the h-likelihood assumption that the survival times within a subject are independent given the random effect we have

$$f(t_i|u_i,\theta) = \prod_{j=1}^{m_i} f(t_{ij}|u_i,\theta) \quad (8)$$

whence, after marginalizing over u and assuming non-informative censoring, (7) becomes:

$$L(\theta,\sigma^2) = \prod_{i=1}^{n}\int_0^\infty g(u_i|\sigma^2)\prod_{j=1}^{m_i}[\lambda(t_{ij}|u_i,\theta)]^{\delta_i}S(t_{ij}|u_i,\theta)du_i \quad (9)$$

4.1   Bivariate Models

Let us consider the bivariate case with $m_i = 2$ so that there are two survival times measured on each subject. After some algebra, we obtain our two models:

Bivariate Weibull-Gamma model:

$$f_f(t_{i1}, t_{i2} \mid \theta) = \frac{(1 + \sigma^2)(\lambda^\rho \rho e^{x'\beta})^2 \lambda^2 (t_{i1}t_{i2})^{\rho-1}}{(1 + \sigma^2 \lambda^\rho [t_{i1}^\rho + t_{i2}^\rho] e^{x'\beta})^{2 + \frac{1}{\sigma^2}}} \tag{10}$$

and Bivariate CTDL-Gamma model:

$$f_f(t_{i1}, t_{i2} \mid \theta) = \frac{(1 + \sigma^2)\lambda^2 p_{i1} p_{i2}}{\left\{1 - \frac{\lambda \sigma^2}{\alpha} \log_e(g_i^2 q_{i1} q_{i2})\right\}^{2 + \frac{1}{\sigma^2}}} \tag{11}$$

Therefore, contrary to some claims that have been made previously, we have been able to use the marginalization approach to obtain the bivariate CTDL-Gamma model. We should note that models (10) and (11) are proportional to their corresponding univariate forms, and can easily be extended to higher dimensional data.

4.2   Correlated Gamma frailty

In the previous section, we assumed shared frailty when dealing with bivariate survival data. However, this assumption may not always be plausible, and hence we should perhaps prefer each of the two survival times measured on an individual to have its own frailty component associated with it. A case that is of particular interest occurs when the two frailty components follow Gamma distributions which are correlated. A substantial amount of research has been done in this field, mainly by Yashin, e.g. Yashin (1995), especially when dealing with twin data, but the thrust of this work is wholly in relation to PH models. No attention has been given to the case where a non-PH hazard is assumed.

Let the two frailties be constructed as $U_1 = Y_0 + Y_1$ and $U_2 = Y_0 + Y_2$, where Y_i are independent Gamma random variables with parameters (k_i, θ_i), $i = 0, 1, 2$. Let us further suppose that $V[U_1] = \sigma_1^2$, $V[U_2] = \sigma_2^2$ and $corr[U_1, U_2] = \rho_u$. We force U_1 and U_2 to be Gamma distributed by assuming $\theta_0 = \theta_1 = \theta_2$. We also retain the earlier assumption of conditional independence of survival times.

For the CTDL model, we have, after some algebra:

$$S(t_1, t_2) = \int_0^\infty \int_0^\infty \int_0^\infty (g_i q_{i1})^{\frac{\lambda}{\alpha}(y_0 + y_1)} (g_i q_{i2})^{\frac{\lambda}{\alpha}(y_0 + y_2)} g(y_0) g(y_1) g(y_2) dy_0 dy_1 dy_2 \tag{12}$$

$$= [1 + \theta\Lambda(t_1)]^{-k_1}[1 + \theta\Lambda(t_2)]^{-k_2}[1 + \theta\Lambda(t_1) + \theta\Lambda(t_2)]^{-k_0} \qquad (13)$$

where $g(y_i)$ are probability density functions of the random variables Y_i, $i = 0, 1, 2$, respectively. A similar form is obtained for the Weibull distribution. A simulation study was performed and the results will appear elsewhere.

5 Final Remarks

In this paper we have extended the non-PH based Gamma frailty and its standard PH-based Gamma frailty competitor to bivariate case. The models we obtained when the frailties are correlated are of a more general form than those commonly used, since we do not assume identical distribution of Y_i, $i = 0, 1, 2$.

Our development of multivariate parametric versions of the non-PH frailty model to deal with correlated survival data and our investigation of correlated frailties opens up further interesting avenues of research. The development of this class of models and various non-parametric, finite mixture, competitors is also being pursued.

Key References

Aalen O.O. (1988) Heterogeneity in Survival Analysis. *Statistics in Medicine*, 7, 1121-1137 .

Ha, I. D., Lee, Y. and Song, J.-K. (2001) Hierarchical likelihood approach for frailty models. *Biometrika*, 88, 233-243.

Hougaard, P. (1984). Life table methods for heterogeneous populations : Distributions describing the heterogeneity. *Biometrika*, 71, 75-83.

MacKenzie, G. (1996) Regression models for survival data: the generalised time dependent logistic family. *JRSS Series D*, 45, 21-34.

MacKenzie, G. (1997) On a non-proportional hazards regression model for repeated medical random counts. *Statistics in Medicine*, 16, 1831-1843.

Yashin, A. I. et al (1995) Correlated Individual Frailty: An Advantageous Approach to Survival Analysis of Bivariate Data. *Mathematical Population Studies*, 5, 145-159.

Assessing Reliability and Agreement of Repeated Measurements by Hierarchical Modeling

Alessandra R. Brazzale[1], Alberto Salvan[2] and Marta Parazzini[13]

[1] Istituto di Ingegneria Biomedica, Consiglio Nazionale delle Ricerche, Italy.
   `alessandra.brazzale@isib.cnr.it`
[2] Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", Consiglio Nazionale delle Ricerche, Italy.
   `salvan@iasi.cnr.it`
[3] Dipartimento di Bioingegneria, Politecnico di Milano, Italy.
   `Marta.Parazzini@polimi.it`

Abstract: We present the use of linear hierarchical models to assess the repeatability and agreement of two or more measurement devices. The idea is illustrated by means of two sets of data. The first considers eight different protocols for the recording of distortion product otoacoustic emissions in Sprague-Dawley rats. The second data set was obtained from the calibration of two types of extremely low frequency magnetic field dosimeters.

1 Background and motivation

At least two concerns must be raised whenever several devices and/or different equipments are used in one and the same study to measure the quantities of interest. The first question regards the reliability of the instruments, that is, whether the reported values reflect the target value being measured. The second point which should be addressed is whether the measurement devices agree, that is, whether they provide under the same experimental conditions measures that may be treated alike. The precision of a measurement device is usually reported in terms of the repeatability standard deviation, while the common measure of agreement in method comparison studies is the intra-class correlation coefficient. In this paper we consider estimates of both obtained under experimental laboratory conditions.

In its original formulation, the model used to represent the calibration data is a one-way random effects model. This formulation rests upon an experimental setup where repeated measures of the same item are taken with

the different methods. Nowadays, laboratories specialized in reliability and method comparison studies tend to adopt more complex calibration procedures. The aim of this paper is to show how linear hierarchical models (Goldstein, 1995) represent a natural extension of the classical approach which allows the experimenter to cope with more elaborate experimental setups. We will illustrated this by means of two data sets provided by two studies who respectively focus on the effects of microwave electro-magnetic fields on hearing and try to assess whether extremely low frequency magnetic fields represent a risk factor for childhood leukemia.

2   Reliability study[1]

2.1   The DPOAE recording data

The DPOAE recording data set contains the distortion product otoacoustic emission (DPOAE) levels recorded from 10 male Sprague-Dawley rats following eight different protocols. Each protocol considers five different frequencies of the stimulating pure tones at which the DPOAEs are measured. The objectives of the study were two-fold: first, to assess the repeatability of the eight protocols used, and, second, to infer the frequency at which the maximum response level is achieved. The animals were tested three times and on both ears separately.

2.2   Model and results

As suggested by the exploratory analysis of the data, a quadratic relationship between DPOAE level, y_{ijkm}^{P}, and tested frequency, x_{ijkm}, was assumed. Random coefficients were introduced to account for individual differences in the mean response level among animals and between the tested ears. Individual models were fitted to the data available for each of the eight recording conditions. The software used is the R (Ihaka and Gentleman, 1996) library nlme developed by Pinheiro and Bates (2000). The final model validated by the data is

$$y_{ijkm}^{P} = (\alpha^{P} + a_{jk}^{P}) + (\beta^{P} + b_{j}^{P})x_{ijkm} + \gamma^{P}x_{ijkm}^{2} + \sigma_{P}\varepsilon_{ijkm}^{P}.$$

Here, the indexes P, i and m identify respectively a particular protocol, the tested frequency and the recording session, b_{j}^{P} is a centered Gaussian random coefficient associated with rat j, and a_{jk}^{P} represents a centered Gaussian random effect that accounts for the difference between the two ears of the jth animal. The repeatability standard deviation associated with

TABLE 1. DPOAE recording data — restricted maximum likelihood estimates of the repeatability standard deviations of the eight protocols used.

protocol	P1	P2	P3	P4	P5	P6	P7	P8
$\hat{\sigma}_P$	6.76	5.96	7.19	6.00	5.93	5.70	6.61	7.65

a particular recording condition identifies with the standard deviation σ_P of the error term in the corresponding model. Table 1 lists the restricted maximum likelihood estimates $\hat{\sigma}_P$ obtained for the eight protocols. The frequency at which the maximum DPOAE response is reached, $x^P_{\max} = -(\beta^P + b^P_j)/(2\gamma^P)$, varies among individual rats, but not with respect to the two ears. On the other hand, a similar calculation shows that the right ear generally produces a higher DPOAE than the left ear.

3 Method comparison study[2]

3.1 The EMDEX$^{\text{TM}}$ calibration data

The EMDEX$^{\text{TM}}$ calibration data consist of the periodical calibrations of 40 EMDEX II$^{\text{TM}}$ and EMDEX Lite$^{\text{TM}}$ magnetic field dosimeters used in the SETIL study. The objectives of the analysis were two-fold: to assess the reliability of the two meter types and to evaluate whether the measurements provided agree. The experimental setup considers six different magnetic flux densities. Three measurements are taken at each nominal value, where, in turn, one of the three sensing coils incorporated into the meter is pointed in the direction of the magnetic field vector. At each occasion, the true magnetic flux density is calculated. Of the 40 instruments considered in our analysis, 21 were calibrated four times and 19 five times.

3.2 Model and results

The preliminary analysis of the data indicated that the absolute measurement error d_{ijkm}, defined as the difference between the measured field strength and the generated magnetic flux density, grows linearly with the true density x_{ijkm} of the target field. We hence used a straight line regression to summarize the mean behaviour of the EMDEX$^{\text{TM}}$ meters. The dependence on the remaining design variables was accounted for by allowing the intercept and slope to vary among instrument type and coil orientation. The SAS procedure PROC MIXED (SAS Institute, 2001) was used to fit the model. The final model validated by the data is

$$d_{ijkm} = (\alpha_i + a_{ijk}) + (\beta + b_{ij} + b_{ijk})x_{ijkm} + \sigma_i \varepsilon_{ijkm},$$

TABLE 2. EMDEX$^{\text{TM}}$ calibration data — Predicted relative errors and 95% prediction intervals cross-classified by instrument type and coil orientation.

	coil 1		coil 2		coil 3	
EMDEX II$^{\text{TM}}$	4.5%	[4.1,4.9]	4.4%	[4.0,4.8]	4.6%	[4.2,4.9]
EMDEX Lite$^{\text{TM}}$	2.5%	[1.9,3.0]	5.1%	[4.5,5.6]	−0.7%	[−1.3,−0.3]

where b_{ij} and (a_{ijk}, b_{ijk}) are centered Gaussian random effects, and where the error variance, σ_i^2, $i = 1, 2$, only depends on the factor instrument type, but not on the orientation of the sensing coils. The remaining indexes, j, k and m, respectively represent the coil orientation, the serial number of the instruments, and the calibration session. The interpretation of the fitted model is straightforward. The fixed effects estimates $\hat{\alpha}_1 = -0.015$ (95% CI: $[-0.017, -0.013]$) and $\hat{\alpha}_2 \equiv 0$ represent the systematic error components associated with the two meter types EMDEX II$^{\text{TM}}$ and EMDEX Lite$^{\text{TM}}$. The estimated overall relative measurement error for both dosimeters amounts to $\hat{\beta} = 4.5\%$ (95% CI: $[4.1\%, 4.9\%]$). The individual relative error for the two instrument types depends on the orientation of the sensing coils. Table 2 summarizes the predicted relative measurement errors and the corresponding 95% prediction intervals cross-classified by instrument type and coil orientation.

Acknowledgments

References

Goldstein, H. (1995). *Multilevel Statistical Models*. Halstead Press, New York.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.

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

SAS Institute (2001). *SAS/STAT$^{\circledR}$ Software, Release 8.2*. SAS Institute Inc., Cary, NC, USA.

Daily Volatility Modelling Using Ultra-High Frequency Data

Christian T. Brownlees[1] and Marco J. Lombardi[1]

[1] Dipartimento di Statistica "G. Parenti" Università degli Studi di Firenze, e-mail: ctb@ds.unifi.it, mjl@ds.unifi.it

Abstract: In this paper we make use of intra-daily information in the modelling of daily volatility. More precisely we will build up a GARCH-like specification which includes intra-daily information. The explanatory power of this model will be compared with standard daily GARCH models.

1   Introduction

One of the issues raised by the advent of Ultra-High Frequency Data in the field of Financial Econometrics is how high frequency information can be exploited in the modelling of lower frequency price dynamics, in particular daily volatility.

One of the most well known stylized fact about high frequency data is the "U" (or "reverse J") pattern that can be observed throughout the day in volumes, absolute returns and number of transactions per interval. The economic rationale for this fact could be that the market participants spend the morning in discounting the information accumulated at night and then the afternoon in trying to anticipate the news that will be released after market closure.

It is thus tempting to specify a intra-daily GARCH structure that makes use of this information and takes into account the different impact on the volatility dynamics of the morning and the afternoon returns.

2   An intra-daily GARCH framework

Let us start introducing our model by splitting the daily close-to-close return r_t into the morning (and overnight) r_t^m and the afternoon r_t^a return:

$$r_t = r_t^m + r_t^a. \tag{1}$$

Let us assume, for simplicity's sake, that the returns have no autocorrelation structure. We will focus our attention on the daily conditional variance h_t, for which we will assume a simple Gaussian GARCH (1,1) structure:

$$h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1}, \tag{2}$$

from which follows that the conditional distribution of r_t is Gaussian with zero mean and variance h_t. If we plug (1) in (2) to express the conditional variance in terms of the intra-daily contributions, we get

$$h_t = \omega + \alpha \left(r_{t-1}^{m2} + 2r_{t-1}^m r_{t-1}^a + r_{t-1}^{a2} \right) + \beta h_{t-1}. \tag{3}$$

Note that the above formulation is just an alternative way to write the standard GARCH (1,1) model and makes no use of the intra-daily information. Instead, the GARCH specification

$$h_t = \omega + \alpha_1 r_{t-1}^{m2} + \alpha_{12} r_{t-1}^m r_{t-1}^a + \alpha_2 r_{t-1}^{a2} + \beta h_{t-1} \tag{4}$$

corresponding to (3) if we enforce the constraints

$$\begin{cases} \alpha_1 = \alpha_2 = \alpha \\ \alpha_{12} = 2\alpha \end{cases} \tag{5}$$

exploits the intra-daily information by allowing for a different effect on the conditional variance of both the morning and afternoon squared returns and their covariance. In order to assess the usefulness of such a specification, one has to verify whether the null hypothesis on the constraints (5) can be rejected. Since the models are nested, we can accomplish that with a simple LR test.

3 Empirical findings

In this section, we will concentrate on the estimation of the models (2) and (4) for a set of four blue chips of the NYSE and we will show how the intra-daily information can successfully improve the performance of the model.

The sample period we will consider is January 1994 – December 1997 (1009 daily observations) and the stocks we will focus on are Dupont (DD), General Electric (GE), Johnson & Johnson (JNJ) and J.P. Morgan (JPM). The daily close-to-close returns have then been split, according to (1), in night–morning returns (from 4pm to 12:30am of the following dat, r_t^m) and afternoon returns (from 12:30am till 4pm, r_t^a), thus yielding a series of 2018 alternated returns.

As far as number of transactions per interval is concerned, the stylized fact we mentioned in the introduction is confirmed by the data in our sample. Figure 1 shows the average number of transactions classified by half hour interval of the day from the opening to the closing of the NYSE.
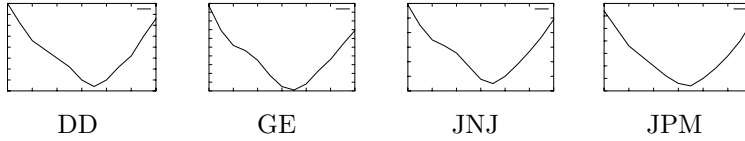
FIGURE 1. Average number of trades classified by half-hour interval of the day from the opening to the closing of the NYSE

First of all, we have applied a standard GARCH (1,1) model to the series of the daily returns. Results are displayed in Table 3.

TABLE 1. Daily GARCH (1,1) estimates

	DD	GE	JNJ	JPM
μ	0.1206	0.1050	0.1071	0.0511
	(2.8190)	(2.5563)	(2.4467)	(1.2956)
ω	0.5783	0.1305	0.1344	0.0433
	(4.4399)	(3.0377)	(2.6533)	(2.1757)
α	0.1986	0.0806	0.0658	0.0495
	(5.2387)	(4.4549)	(4.6778)	(4.1205)
β	0.5583	0.8520	0.8705	0.9285
	(7.2355)	(25.923)	(27.539)	(46.919)
Log-L	-1824.915	-1724.936	-1782.788	-1711.035

As we anticipated, daily returns were then split and the model of equation (4) was applied to the double-length series; results are presented in Table 3.
If we consider the cross-correlation coefficient α_{12}, we observe that in this case, albeit having a positive sign, it is not close to what we should expect according to the constraints (5), that is $\alpha_{12} = 2\alpha$. The fact that the coefficient is positive indicates that if we observe two returns of different signs in the morning and the afternoon of the previous day, we should expect a negative impact on volatility. We could argue that two returns of different signs (provided they are of small magnitude) are symptomatic of market in as state of equilibrium, so that operators tend not to modify their positions. The GARCH coefficients $\hat{\beta}$ are of the same order of magnitude as their daily counterparts. However, it could be argued that the intra-daily β should be smaller than its daily counterpart because, given that an intra-daily specification exploits a larger amount of information, there should remain less unexplained patterns to be captured by the β coefficient. We have thus performed an asymptotic z-test on the difference of the two parameters, with the null hypothesis $\beta_D = \beta_{ID}$ against the alternative $\beta_D > \beta_{ID}$.

TABLE 2. Infradaily GARCH estimates

	DD	GE	JNJ	JPM
μ	0.0580	0.0564	0.0689	0.0401
	(2.9318)	(2.8977)	(3.2236)	(2.5542)
ω	0.2813	0.0770	0.1259	0.0506
	(6.9310)	(4.1725)	(4.2392)	(25.3379)
α_1	0.0885	0.0953	0.0504	0.0713
	(4.9236)	(6.2946)	(4.5133)	(13.4495)
α_2	0.1848	0.0527	0.0756	-0.0649
	(6.3342)	(3.5675)	(4.6778)	(-20.106)
α_{12}	0.1347	0.0341	0.0519	0.0285
	(4.0351)	(1.6929)	(2.5394)	(8.2063)
β	0.4993	0.7674	0.7629	0.8919
	(8.7724)	(21.921)	(20.902)	(192.683)
Log-L	-1053.796	-836.1532	-1030.566	-1042.176
LR	413.628	427.311	352.107	95.513
AIC_{ID}	1.0514	0.8355	1.0283	1.0399
AIC_D	1.2546	1.0454	1.2010	1.0853
$\beta_D = \beta_{ID}$	-21.529	-65.289	-83.707	-57.947

Indeed results, reported in the last row of Table 3, indicate that the null hypothesis is always rejected.

Finally, we have performed a LR test to verify whether the constraints (5) can be assumed to be consistent with the data, thus leading to the conclusion that the intra-daily The outcome of the test clearly indicates the superior explanatory capability of our model. This is confirmed by the comparison of the AIC's which is reported in the table.

A simple extension of the News Impact Curves (NIC) (Engle and Ng, 1993) can be exploited as an appealing device to visualize the impact of the morning and afternoon returns on the daily volatility. The NIC curve shows the impact on volatility as a function of the daily return, in a given ARCH-like model framework. Analogously, given our intra-daily GARCH specification, we will construct a News Impact Surface (NIS) which will be used to visualize the joint impact of the intra-daily returns on volatility. The News Impact Surface (NIS) in our GARCH framework is expressed by the following expression:

$$NIS = \alpha_0 + \alpha_1 r^{m2} + \alpha_2 r^{a2} + \alpha_{12} r^m r^a,$$

where $\alpha_0 = \omega + \beta \sigma^2$.

The pictures shown in Figure 2 are 3D plots of the NIS function, given the model estimates of our sample of tickers in Table 3.

An easy and interesting way to interpret these graphs is to analyze theirs

DD   GE
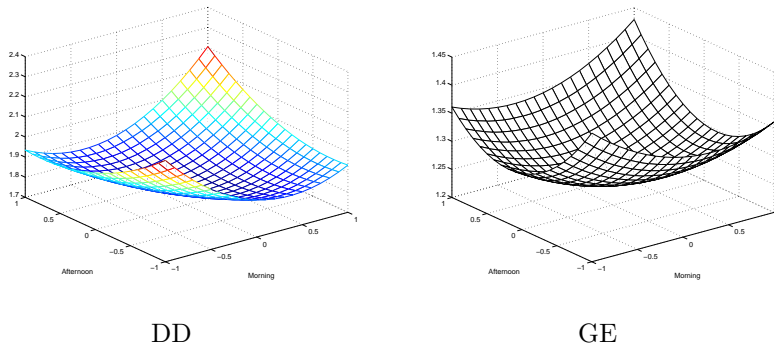
FIGURE 2. News Impact Surfaces

sections. For a given level of the morning or afternoon return, the corresponding profile of the NIS can be considered as a NIC. First of all, the presence of the cross-correlation coefficient makes the NIC's asymmetric; the fact that it is always positive implies a higher impact on volatility when the two returns have the same sign. The steepness of the NIS only depends on the magnitudes of α_1 and α_2, whereas its concavity depends on their signs. The plots display different patterns of steepness and concavity.

4 Conclusions

We have introduced a new intra-daily GARCH specifications that allows for a different impact on volatility of the morning and afternoon returns. This model is consistent with the stylized facts of intra-daily pattern of market activity, which tend to decrease during the morning and increase in the afternoon. The empirical application we have presented points out that the proposed model performs well and appropriately exploits the intra-daily information.

References

Engle, R.F. and Ng, V.K. (1993) Measuring and testing the impact of news on volatility, *Journal of Finance*, **48**, 1749–1778.

Andersen, T. G. and Bollerslev, T. (1997) Intraday Periodicity and Volatility Persistence in Financial Markets, *Journal of Empirical Finance*, **4**, 115–158.

Control of the False Discovery Rate with Bayes Factors. An application to Microarray data analysis

S. Cabras[1] and W. Racugno[2]

[1] Department of Statistics, University of Florence. V.le Morgagni, 59 - 50134 Firenze (*Italy*). `cabras@ds.unifi.it`
[2] Department of Mathematics, University of Cagliari (*Italy*)

Abstract: In this work we consider Multiple Hypothesis Testing approach to gene expression data analysis. We focus on controlling the False Discovery Rate by using Bayes Factors as test statistics and approximating their sampling distribution under the null hypothesis.

Keywords: Intrinsic Bayes factors; Multiple hypothesis testing; Rejection region.

1 Introduction

Microarrays are emerging as a powerful and cost-effective experiments for large scale analysis of gene expression. These experiments are typically done in a case-control study framework where thousands of genes are simultaneously compared in order to *discover* which are differentially expressed. The statistical approach to data analysis is typically based on *Multiple Hypothesis Testing* (MHT), because we have to account for the multiplicity arising when testing m *null hypotheses* $H_i^0 = \{$gene i is not differentially expressed$\}$ for $i = 1, \ldots, m$, and m is of order of thousands. In microarray data analysis, MHT is mainly concerned in controlling the *False Discovery Rate* (FDR) which is the rate of false rejections (discoveries) among all rejections (see Storey 2003 and Benjamini and Hochberg 1995 for a review and bibliography on FDR). The main problem with MHT is to construct a rejection region Γ_α for a single test i and to calculate the type I error α corresponding to Γ. For a chosen test statistic T_i we have to calculate $\Pr^{F_0(t)}\left(T \in \Gamma_\alpha | H_i^0\right)$ where $F_0(t)$ represents the sampling distribution of T_i under the null hypothesis (in the sequel *null distribution*). It is usual to consider as test statistics the set of m ordered p-values with $F_0(p)$ the $Uniform(0,1)$ distribution. The rejection region for each test takes the form of $\Gamma_{p_i} \doteq (0, p_i)$ with $p_i = \Pr^{F_0(p)}\left(P \leq p_i | H_i^0\right)$ which leads to the frequentist interpretation of the p-values (*frequentist p-values*). Unfortunately this interpretation does not generally hold in particular when H_i^0 is a composite hypothesis (or model) and no ancillary statistics are available

for the involved nuisance parameters. In this case $F_0(p)$ depends on the way we eliminate the nuisance parameters (see Bayarri and Berger, 2000). On this purpose we investigate the use of the Bayes Factor (BF) B_i as test statistic T_i, where $B_i = m_1(\mathbf{x}_i, \mathbf{y}_i)/m_0(\mathbf{x}_i, \mathbf{y}_i)$ is the ratio of the marginal distributions $m_j(\mathbf{x}_i, \mathbf{y}_i)$ under the hypothesis H_i^j for $j = 0, 1$ for independent single gene expression measurements $\mathbf{x}_i = (x_1^i, \ldots, x_n^i)$, $\mathbf{y}_i = (y_1^i, \ldots, y_n^i)$. As the use of p-values does not require the indication of the alternative hypothesis (or model) H_i^1, we will compare BFs and p-values in those cases where the model classes for $m_j(\mathbf{x}_i, \mathbf{y}_i)$ are known. For this reason we recommend the use of BF after the model checking phase. We argue that a good reason to use BF instead of the p-values is that under H_i^1, $B_i \to \infty$ as $n \to \infty$ while p_i is still random distributed in $(0, 1)$. In gene expression data analysis difficulties of elicitation on model parameters make the use of non-informative priors unavoidable. This leads to well known problems in determining the BF because the adopted prior distributions are often improper. These difficulties are dealt with the *intrinsic BF*, the *fractional BF* and their modifications such as the *intrinsic procedures*. For a review and bibliography on BFs with improper priors see Moreno, Bertolino and Racugno, (1998-1999) and references therein. We consider the set of random rejection regions $\Gamma_{\alpha_i} \doteq (b_i, \infty)$ where b_i is an observed B_i and we approximate $\alpha_i = \mathrm{Pr}^{F_0(b)}(B_i > b_i | H_i^0)$ using a Monte Carlo sum by simulating B_i under H_i^0. In this way the FDR can be estimated on the set of ordered α_i which can be viewed as the analogues to p-values, but calculated on the null distribution of BFs. The null distribution of BF has not been considered as orthodox in the Bayesian paradigm, because it supposes the use of BFs which have never been observed. However, the way some authors proposed to summarize the evidence arising from BF are an attempt to calibrate the BF with respect to categories which do not formally arise from experimental data (see for instance Kass and Raftery (1995) and references therein). In this case our categories are the α_is which have a meaning for those procedures that estimate or control the FDR. Section 2 contains the BFs we use for Normal and Gamma models. Section 3 presents a simulation study to show the potential of the procedure and an application to a data set from a controlled experiment. Conclusions and further remarks are contained in Section 4.

2   Bayes Factors for Normal and Gamma models.

The problem is usually to test the equality of unknown means of gene expressions in the case $\mathbf{X}_i$ and in the control $\mathbf{Y}_i$. As we assume the same statistical model on each gene so we will suppress the subscript i unless necessary. We consider here, as an exemplification, the Normal and the Gamma model which are often assumed in microarray data analysis. The Normal model is assumed after *Normalization Process* of the data (see

Dudoit *et al.* 2001 for bibliography), while the Gamma model is assumed to analyze the outcome from cDNA experiments because it easily allows to control the common variation coefficients of gene expression measurements $\mathbf{X}_i$ and $\mathbf{Y}_i$ (Newton *et al.* 2002).

The Normal model. We restate the hypothesis testing in terms of model selection by comparing $\mathcal{M}_0 : f_0\left(\mathbf{x},\mathbf{y}|\theta_0\right) = N\left(\mu,\sigma_X^2\right) N\left(\mu,\sigma_Y^2\right)$, $\pi_0^N\left(\theta_0\right) = c_0\left(\sigma_X\sigma_Y\right)^{-1}$ versus $\mathcal{M}_1 : f_1\left(\mathbf{x},\mathbf{y}|\theta_1\right) = N\left(\mu_X,\sigma_X^2\right) N\left(\mu_Y,\sigma_Y^2\right)$, $\pi_1^N\left(\theta_1\right) = c_1\left(\sigma_X\sigma_Y\right)^{-1}$, where $\theta_0 = (\mu,\sigma_X,\sigma_Y)$, $\theta_1 = (\mu_X,\mu_Y,\sigma_X,\sigma_Y)$ and π_0^N, π_1^N denote the assumed reference priors of θ_0 and θ_1 respectively, with c_0 and c_1 arbitrary constants. Testing the equality of the means in two Normal populations is a time honored problem in statistics, in particular as it is well known when $\sigma_X = \sigma_Y$ the test corresponds to the t-test, otherwise the Behrens-Fisher problem arises. If the samples are paired (as in cDNA experiments) and $\sigma_X^2 = \sigma_Y^2$ the hypothesis testing can be viewed as the test on the mean of differences $\mathbf{d}_i = (\mathbf{x}_i - \mathbf{y}_i)$. This test is often used to check the zero mean of M_i coordinates in a MA-plot after data Normalization (Dudoit *et al.,* 2001). In this particular case the test becomes $H_0 : \mathbf{d} \sim N\left(0,\sigma_d^2\right)$ against $H_1 : \mathbf{d} \sim N\left(\mu,\sigma_d^2\right)$ using the opportune reference priors. For the test of $\mathbf{d}_i$ we compare the p-values arising from t-test against the limiting intrinsic BF, B^{Lim}, the fractional BF, B^F and the Schwarz approximation, B^S. The derivation of B^{Lim}, B^F and B^S is contained in Moreno, Bertolino, Racugno (1998). For simplicity in the Behrens-Fisher problem we compare the t-test with Welch correction, p^{Welch} only against the BF obtained with the Schwarz approximation, $B^{BF(S)}$ whose expression is contained in Moreno, Bertolino and Racugno (1999). We will mainly consider the test of $\mathbf{d}_i$ and the Behrens-Fisher problem.

The Gamma model. The model selection is between $\mathcal{M}_0 : f_0\left(\mathbf{x},\mathbf{y}|\theta_0\right) = Gamma\left(a,\theta\right) Gamma\left(a,\theta\right)$, $\pi_0^N\left(\theta_0\right) = c_0\theta^{-1}\vartheta\left(a\right)$ and $\mathcal{M}_1 : f_1\left(\mathbf{x},\mathbf{y}|\theta_1\right) = Gamma\left(a,\theta_X\right) Gamma\left(a,\theta_Y\right)$, $\pi_1^N\left(\theta_1\right) = c_1\left(\theta_X\theta_Y\right)^{-1}\vartheta\left(a\right)$, where $\vartheta\left(a\right) = \sqrt{\psi^{(1)}\left(a\right) - a^{-1}}$ and $\psi^{(1)}\left(a\right)$ is the trigamma function. $\pi_0^N\left(\theta_0\right)$ is the reference prior and $\pi_1^N\left(\theta_1\right)$ has been assigned without changing the prior for a according to Kass and Wasserman, (1996). In this case the BF is not available in closed form and we will approximate it via Markov Chain Monte Carlo. This solution leads to time consuming simulations and therefore we will consider only fractional BF, $B_{\Gamma}^F = B_{10}^N\left(\mathbf{x},\mathbf{y}\right) B_{01}^b\left(\mathbf{x},\mathbf{y}\right)$ with $b = 3/n$ where

$$B_{10}^N\left(\mathbf{x},\mathbf{y}\right) = \frac{\int_{\Re^+} \frac{p^a}{\Gamma(a)^{2n}}\left[\Gamma\left(na\right)\right]^2 \left(s_x s_y\right)^{-na} \vartheta\left(a\right) da}{\int_{\Re^+} \frac{p^a}{\Gamma(a)^{2n}}\Gamma\left(2na\right) s^{-2na}\vartheta\left(a\right) da} \qquad (1)$$

$$B_{01}^b\left(\mathbf{x},\mathbf{y}\right) = \frac{\int_{\Re^+} \Gamma\left(6a\right) p^{3a/n}\Gamma\left(a\right)^{-6}\left(s\right)^{-6a}\vartheta\left(a\right) da}{\int_{\Re^+} \left[\Gamma\left(3a\right)\right]^2 p^{3a/n}\Gamma\left(a\right)^{-6}\left(s_x s_y\right)^{-3a}\vartheta\left(a\right) da} \qquad (2)$$

with $p = \prod_{i=1}^{n} x_i y_i$, $s_x = \sum_{i=1}^{n} x_i$, $s_y = \sum_{i=1}^{n} y_i$ and $s = s_x + s_y$. Let $\varphi\left(a\right)$ represents the kernel of the distributions in a appearing in the integrals

(1) and (2), we approximate each distribution with a Metropolis-Hasting algorithm using as proposal a Gamma distribution with median equal to $\overline{a} = \max_{a \in R^+} \varphi(a)$ and variance equal to $[\varphi''(\overline{a})]^{-2}$. We compare B_Γ^F with the p-value obtained by using the test statistic $t(\mathbf{x}_i, \mathbf{y}_i) = \overline{x}_i / \overline{y}_i$ whose null distribution is a Multiple Scale Beta of the II kind with shape parameter a replaced by its maximum likelihood estimator $\widehat{a}$. We call this the *plug-in* p-value, p_{plug}, which approximates conservatively the type I error (Bayarri and Berger, 2000).

3   Simulation study and application

We numerically investigate the behavior of mentioned BFs and p-values by simulating J balanced microarray experiments with n replications. We consider experiments of m genes where $m' \leq m$ have different means with respect to the others genes $m - m'$ genes. For each simulated experiment we order all genes according to the p_i and α_i adjusted using the Benjamini and Hochberg procedure (1995), the q-values (Storey, 2003) and the Bonferroni correction, which control different error measurements in MHT. We finally collect the rank assigned to the m' genes and look at the distribution of ranks across $J = 100$ simulations. The more the ranks are concentrated around 1 the more we are likely to detect the m' genes as differentially expressed. When testing the means of $\mathbf{d}_i$ with $m' = 5$, $m = 1000$ we obtain that the p-value from classical t-test and BFs B^{Lim}, B^F, B^S lead to the same results with a sample size $n \leq 20$, but the distributions of the ranks for BFs are more concentrated around 1 for larger sample sizes. For the comparison of B^F versus p^{Welch} with $\sigma_X = K\sigma_Y$, $K = 2$ we find that the ranks assigned using B^F are much more concentrated around 1 for $n \geq 4$. We argue that this is due to the fact that in the Behrens-Fisher problem the null distribution of B^F is more robust to K than the p-value with the Welch correction. This argument applies in particular to the comparison of B_Γ^F versus p_{plug} for the Gamma model where B_Γ^F leads to smaller ranks than p_{plug} with $n \geq 4$.

We consider the application of the mentioned BFs to the `eset12` data set available at `www.biocondutor.org`. Data come from 24 HGU95a Affy chips ($n = 12$ replications) each containing 12626 genes with, 16 genes spiked at different concentrations in two populations under comparison. For the Gamma model we consider subsets of $n \geq 4$ replications and we obtain that only B_Γ^F allows to detect the 16 genes as differentially expressed at $FDR < 0.05$. This does not happen neither with the p_{plug} nor with the other BFs for the Normal model even using all data.

4   Conclusions

The main problem in using BF is the computational effort because of the need to compute a BF for each gene under test. This problem becomes

important in particular when BFs are not available in closed form such as the case for the Gamma model. Nonetheless we conclude that BF are useful test statistics in MHT for controlling the FDR especially when the considered models, such as the Gamma, do not allow to use other ancillary test statistics more than the BF. In fact differences in the performance between p-values and BF are evident at small sample sizes for the Gamma model and in the Beherens-Fisher problem, while when testing the means of $\mathbf{d}_i$ differences between BF and p-values are not so evident unless a very large sample size is used. In this latter case the simulations results are in favor of BF. These results are relevant in microarray data analysis because the sample size is usually small, due to the prohibitive costs for each replication, and when we cannot assume normality.

References

Bayarri, M. J. and Berger, J. O. (2000). p-values for Composite Null Models. *Journal of the American Statistical Association*, **95**, 1127-1142.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society B*, **57**, 289-300.

Dudoit, S.; Yang, Y. H., Callow, M. J. and Speed, T. P. (2001). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111-139.

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

Kass, R. E. and Wasserman, L. (1996). The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association*, **91**, 1343-1370.

Moreno, E., Bertolino, F. and Racugno, W. (1998). An intrinsic limiting procedure for model selection and hypothesis testing. *Journal of the American Statistical Association*, **93**, 1451-1460.

Moreno, E., Bertolino, F. and Racugno, W. (1999). Default Bayesian analysis of the Behrens-Fisher problem. *Journal of statistical planning and inference*, **81**, 323-333.

Newton, M.A., Kendziorski, C. M., Richmond, C. S., Blattener, F. R. and Tsui, K. W. (2002). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, **8**, 37-52

Storey, J. D. (2003) The Positive False Discovery Rate: A Bayesian Interpretation and the q-value. *Annals of Statistics*, **31**, 2013-2035.

Parametric vs semiparametric in interval censored data

G. Parrinello[1] , S. Calza[1] , U. Valentini[2], A. Cimino[2], A. Decarli[1]

[1] Section of Medical Statistics and Biometry, University of Brescia, Italy e-mail parrinel@med.unibs.it
[2] Diabetes Unit Spedali Civili Brescia, Italy

Keywords: interval censored, parametric survival model, cox model, diabetes nephropathy.

1 Background

Type II Diabetes Mellitus (T2DM) is one of the most common endocrine diseases in all populations. Consequently, the knowledge of the factors influencing the incidence of T2DM-related complications is very important.

2 Population

An observational study including a cohort of 1,810 T2DM patients attending the Diabetic Unit (DU) at the Spedali Civili in Brescia, from January 1990 to October 1997.

3 Methods

T2DM is a systemic pathology, so that it is not possible to analyse the prognostic factors related to Incipient Diabetic Nephropathy (IDN) without taking into account the associated complications. A survival model for competing risks permits to overcome this problem. Recently there has been an increasing use of the Cox model adapted for the presence of competing risk, while less attention has been given to parametric models for competing risks, particularly in the area of T2DM studies. For this purpose we have adapted the Lunn-McNeils approach for the analysis of competing risk in the Cox model to the parametric survival regression models, using different distributions suggested by JK Lindsey. The models we propose have to take into account that the time of onset of IDN is heavily interval censored, as the assessment of IDN is based on blood samples. Therefore the exact time of onset is unknown. For this reason we compared the results from a

parametric regression for interval-censored data, with those obtained using the mid-point of the interval or those obtained using the right extreme of the interval of censoring. Parametric models seem to smooth naturally the data using adjacent "information" and are less influenced from interval-censoring.

References

Lunn M. and McNeil D. (1995) Applying Cox Regression to Competing Risks. *Biometrics*, 51524-532.

Parrinello G. and oth. (2001) Analysis Of Incipient Diabetic Nephropathy In Type Ii Diabetes Mellitus Patients Through The Application Of A Parametric Model For Survival Data With Competing Risks - 22st ISCB Meeting, Stockholm,Proceedings

Lindsey, J.K. (1998) A study of interval censoring in parametric regression models. *Lifetime Data Analysis*, **4**, 329-354.

Regression models for the analysis of psychiatric data

Luisa Canal[1], Rocco Micciolo[1]

[1] University of Trento

Abstract: Three regression models were fitted to a set of psychiatric contacts: two parametric (based on the negative binomial distribution and on the generalised Waring distribution (GWR)) and one semiparametric (based on the cumulative mean function of the number of contacts). They were all able to account for the large amount of overdispersion and gave similar estimates of the regression coefficients and of their SEs. However, the GWR model could give additional information to the clinician owing to the possibility to quantificate the *proneness* and the *liability* of different sets of patients in contacting psychiatric services.

Keywords: Negative binomial regression, Semiparametric regression, Psychiatric contacts, Recurrent events, Waring distribution.

1 Introduction

Psychiatric data collected in a psychiatric case register consist of a series of contacts made by subjects within the psychiatric agencies of a selected geographic area. When these data are analysed, one of the most striking features is represented by a large amount of overdispersion. In particular, the distribution of the number of psychiatric contacts shows often a large number of zeroes and a very long right tail. In a previous study Canal & Micciolo (1999) found that the generalised Waring distribution fits well the observed frequencies. Moreover, the variance of this distribution can be divided in three components, named *liability*, *proneness* and *random*. In accident theory, differences in exposure to external risk of accident from person to person are known as differences in accident *liability* as distinguished from constitutional or internal differences which are known as differences in *proneness*. In a psychiatric context, *liability* and *proneness* could be considered as due to exogenous and to endogenous factors. Effects of proneness and liability are confounded when the negative binomial is employed. In this study a regression model based on the generalised Waring distribution will be presented and the results obtained on a data set of psychiatric contacts coming from the South-Verona Psychiatric Case Register (Tansella, 1991) will be compared with those found employing the negative binomial regression (Lawless, 1987; Long, 1997) and a semiparametric regression based on the Mean Function (Lawless & Nadeau, 1995).

2   Regression models

The probability function of the generalised Waring distribution is:

$$p(y) = \frac{\Gamma(\rho + a)\Gamma(\rho + k)}{\Gamma(\rho)\Gamma(\rho + a + k)} \times \frac{a_{[y]}k_{[y]}}{(\rho + a + k)_{[y]}} \times \frac{1}{y!}, \qquad (1)$$

where the parameters ρ, a, k must be all greater than zero. Since the expected value is $\frac{ak}{\rho - 1}$ and the variance is $\frac{ak(\rho + k - 1)(\rho + a - 1)}{(\rho - 1)^2(\rho - 2)}$, ρ must also be greater than 2. The *Pochhammer symbol* $a_{[y]}$ is defined as $a_{[y]} = a(a + 1)\cdots(a + y - 1)$. Let $\mathbf{x_i}$ be a $(p + 1)$ vector of p covariates plus a constant for the intercept term associated with the individual i and assume that

$$E[Y_i|\mathbf{x_i}] = \frac{ak}{\exp(-\mathbf{b}'\mathbf{x_i})} \qquad (2)$$

where $\mathbf{b}$ is a $(p + 1)$ vector of regression parameters and Y_i are mutually independent random variables following a generalised Waring distribution with parameters $a, k, \rho_i = 1 + \exp(-\mathbf{b}'\mathbf{x_i})$. If there is only one dummy covariate, then $\rho_i = 1 + \exp(b_0 + b_1 x_i)$ and $\frac{E[Y_i|x_i=1]}{E[Y_i|x_i=0]} = \exp(b_1)$.

Two other regression models were fitted to the same data set. The first was the negative binomial regression (Lawless, 1987; Long, 1997):

$$Pr[Y = y_i|\mathbf{x_i}] = \frac{\Gamma(s + y_i)}{\Gamma(s)\Gamma(y_i + 1)} \left(\frac{s}{s + m_i}\right)^s \left(\frac{m_i}{s + m_i}\right)^{m_i}, \qquad (3)$$

where $E[Y_i|\mathbf{x_i}] = m_i = (\exp \mathbf{b}'\mathbf{x_i})$ and s is a shape parameter (the reciprocal of s is sometimes referred to as the overdispersion parameter).

The second, which takes into account also the precise event times, was a semiparametric model (Lawless & Nadeau, 1995) based on the Cumulative Mean Function of the number of events $N(t)$ occurring over the interval $[0, t]$: $M(t) = E[N_i(t)]$. This method, which focus on mean functions for processes of recurrent events and do not involve a full probabilistic specification of the processes, is rather widely applicable. The estimator $\hat{M}(t)$ is given by

$$\hat{M}(t) = \sum_{u=0}^{t} \hat{m}(u) \qquad (4)$$

where $\hat{m}(u)$ is the mean number of events observed at time u calculated dividing the total number of events $n(u)$ observed at time u by the number of subjects $\delta(u)$ who are still under observation at time u : $\hat{m}(u) = n(u)/\delta(u)$. A regression model can be set up including the effect of a covariate vector $\mathbf{x_i}$ in a multiplicative way: $m_i(t) = m_0(t) \times \exp(\mathbf{b}'\mathbf{x_i})$; $m_0(t) \geq 0$ is a baseline mean function. In this case $\mathbf{b}$ is a vector of p regression coefficients which does not include an intercept term. The estimating equations for $\mathbf{b}$ together with a robust estimate of its variance can be found in Lawless & Nadeau (1995).

3   Patients and methods

Patients who entered the South-Verona Psychiatric Case Register in the period 1 January 1979 to 31 December 1991 were included in the study. All subjects were followed for 13 weeks. For each patient the total number of contacts in the 91 days of follow-up as well as the day at which each contact was observed were known. The following covariates were also available: gender, occupational status, diagnosis, referral source of the first contact, type of the first contact.

The three regression models described above were fitted to these data. Parameter estimates for the negative binomial regression (NBR) were found employing the procedure NBREG in STATA 7.0 (Stata Corp., 2001). Estimating equations for the regression model based on the mean function (MFR) were solved using Mathematica 4.1 (Wolfram, 1999). Parameter estimates for the generalised Waring regression (GWR) model were obtained by maximum likelihood; for computational purposes, the parameter restrictions $a > 0$, $k > 0$ were incorporated re-parameterising the log-likelihood function so that these constraints were eliminated: the parameters a and k in (1) were replaced by $\exp(a_0)$ and $\exp(k_0)$ respectively. To find the maximum of the observed log-likelihood the algorithm proposed by Morabito and Marubini (1976) was employed; their strategy, which resorts to a combination of the steepest descent and the Newton-Raphson method, is suitable for both speed of convergence and numerical accuracy.

4   Results

A total number of 3454 subjects were included in this study, with a total number of 6913 contacts. Table 1 shows the parameter estimates of the regression coefficients obtained using the three regression models together with the corresponding standard errors. Conclusions in terms of significance tests were quite similar. A significantly higher number of contacts was found for unemployed subjects, for patients with an unplanned first contact and for those with a self-referral (or referred by relatives). As far as diagnosis is concerned, higher contacts were found for schizophrenic patients.

The component of variance of the GWR attributable to liability ranged from 5.2% (for other diagnosis) to 8.6% (for self-referred subjects) and to proneness from 68.5% (for other diagnosis) to 90.4% (for patients with a diagnosis of schizophrenia); affective disorders, organic psychoses, alcoholism and personality disordersshowed a proneness between 80% and 90%.

5   Conclusions

The estimates of the regression coefficients obtained using three different approaches were substantially similar. Since for both the GWR and the

TABLE 1. Estimates of the regression coefficients for the generalised Waring regression (GWR), the negative binomial regression (NBR) and the mean function regression (MFR).

VARIABLES	ESTIMATES			STANDARD ERRORS		
	GWR	NBR	MFR	GWR	NBR	MFR
Gender						
Females vs Males	-0.104	-0.090	-0.090	0.063	0.057	0.063
Occupational status						
Unempl. vs Empl.	0.811	0.758	0.758	0.134	0.109	0.110
Other vs Empl.	0.090	0.106	0.106	0.063	0.059	0.065
Diagnosis						
Affective Dis. vs Schiz.	-0.977	-0.892	-0.892	0.142	0.111	0.101
Organic Psych. vs Schiz.	-0.816	-0.699	-0.699	0.217	0.186	0.206
Alc. / pers. dis. vs Schiz.	-0.872	-0.745	-0.745	0.153	0.121	0.124
Neurotic Dis. vs Schiz.	-1.316	-1.210	-1.210	0.142	0.111	0.105
Other Dis. vs Schiz.	-1.459	-1.348	-1.348	0.146	0.115	0.111
Referral source						
GPs vs Self-referral	-0.240	-0.265	-0.265	0.102	0.091	0.085
Others vs Self-referral	-0.556	-0.510	-0.510	0.068	0.060	0.068
First contact						
Unplanned vs Planned	0.710	0.683	0.683	0.070	0.062	0.066

NBR the precise event times were not considered, it appears that, to assess the covariate effects, the total number of contacts during the study period contains much of the information about **b**. Also standard errors were quite similar for the three models. Unlike those obtained for the GWR and for the NBR, the variance estimates for the MFR were robust moment-based and valid quite generally (Lawless & Nadeau, 1995). The key assumption, that is that the end of observation times be independent of the event process, is likely fulfilled in our study, since it was fixed in advance for all subjects. Since (i) the follow-up times were all equals (so that the same number of subjects was at risk at each time), (ii) only dummy variables were used as covariates and (iii) only "univariate" analyses were performed, an exact solution was obtained for the estimating equations of the regression coefficients of the MFR model; for a categorical variable with k levels, coded with $k-1$ dummies, the estimate of the j-th regression coefficient is $\ln\left[n_0 N_j / (n_j N_0)\right]$, where n_j is the number of subjects in the category $j+1$ and N_j is the overall number of contacts of the subjects in the category $j+1$ (the deponent 0 indicates the reference category).

It is worth noting that the estimates for the MFR model and those obtained from the NBREG procedure for the NBR were the same, at least within the numerical accuracy of the STATA output. So it appears that a semi-parametric model and a parametric model gave the same estimates as far

as the regression coefficients are concerned; however this is true only if univariate analyses are performed or, in case of multivariate analyses, if saturated models are fitted.

As far as results obtained using the GWR are concerned, we think that this model, despite the similar results, the higher number of parameters to be estimated and a more heavy computational job, could give additional useful information to the clinician owing to the possibility to divide the total variance in three components. Since the generalised Waring distribution is symmetrical in a, k, the proneness and the liability component cannot be universally identified. However since in our case one of the variance component was much larger than the other, we think to be justified in attributing this component to proneness. In the data set analysed, endogenous factors appear to be quite important and account for 70% (or more) of the variability, while the percentage of variance due to exogenous factors is similar for all the categories of patients (between 6% and 8%). Endogenous factors appear to be more important for patients with a diagnosis of schizophrenia, unemployed, self-referred and with an unplanned first contact.

In conclusion, even if for comparison purposes between categories of patients, any one of the selected regression models can be employed, we think that the GWR could be quite useful in comparing psychiatric data coming from different geographic settings covered by a psychiatric case register.

References

Canal, L., and Micciolo, R. (1999). Modelli probabilistici per l'analisi dei contatti psichiatrici. *Epidemiologia e Psichiatria Sociale*, **8**, 47-55.

Lawless, J.F. (1987). Regression methods for Poisson process data. *Journal of the American Statistical Association*, **82**, 808-815.

Lawless, J.F., and Nadeau, C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics*, **37**, 158-168.

Morabito, A., Marubini, E. (1976). A computer program suitable for fitting linear models when the dependent variable is dichotomous, polichotomous or censored survival and non-linear models when the dependent variable is quantitative. *Computer Programs in Biomedicine*, **5**, 283-295.

STATACORP, (2001). *Stata Statistical Software: Release 7.0*. College Station, TX: Stata Corporation.

Tansella, M. (ed) (1991). Community-based psychiatry: long-term patterns of care in South-Verona. *Psychological Medicine Monograph*, S. 19.

Wolfram, S. (1999). *The Mathematica Book,* 4th ed. Wolfram Media / Cambridge University Press, Cambridge.

A statistical method for the estimation of childhood cancer prevalence among adults

A. Gigli[1], A. Simonetti[1,2], R. Capocaccia[3] and A. Mariotto[4]

[1] Institute for Research on Population and Social Policies, CNR, Roma, Italy
[2] Dept. Statistics and Probability, Universita' "La Sapienza", Roma, Italy
[3] National Health Institute, Roma, Italy
[4] National Cancer Institute, Bethesda, USA

Abstract: We present a method to estimate the complete prevalence of patients of current age x who have been diagnosed with childhood cancer in the age interval $(0, t_0)$, based on data observed for L years.

1 Introduction

Estimating the number of individuals in a population that had cancer in their childhood is relevant, because prognosis of many childhood cancers is fairly good, and most young patients become long-term survivors; however, psycological or physical consequences of the disease may persist for their entire life, due to the aggressiveness of the treatments and to the increased risk of subsequent cancers, and they may need extra medical care. Cancer prevalence is defined as the proportion of people alive on a certain date who have been previously diagnosed with the disease; for a fixed birth cohort c it can be formalized as convolution of incidence and survival functions:

$$N_x(0, x) = \int_0^x I(t)S(x - t, t)dt, \qquad (1)$$

where $N_x(0, x)$ is the prevalence at current age x of cases diagnosed between age 0 and x, $I(t)$ is the incidence hazard at age t, $S(x - t, t)$ is the survival function at age x of patients who were diagnosed at age t. In a population covered by cancer registration, where data on diagnosis and life status of all incident cases are collected, prevalence can be estimated by enumerating the number of incident cases that are still alive at a fixed date of prevalence, and correcting for cases lost to follow up. This estimator, called *Limited Duration Prevalence* (LDP), is based on a limited observational period L (from the starting date of registration to the date of prevalence):
$\hat{N}_x(x - L, x)$ is the estimate of prevalence of patients of current age x who were diagnosed in the last L years. To take into account cases diagnosed

before the beginning of the registry, the *Completeness Index*, defined as the fraction of modelled prevalence which is observed, was introduced by Capocaccia and De Angelis (1997):

$$R_x(L; \hat{\psi}) = \frac{N_x(x - L, x; \hat{\psi})}{N_x(0, x; \hat{\psi})} = \frac{\int_{x-L}^x I(t; \hat{\psi}) S(x - t, t; \hat{\psi}) dt}{\int_0^x I(t; \hat{\psi}) S(x - t, t; \hat{\psi}) dt}, \qquad (2)$$

where I and S are parametric functions and $\hat{\psi}$ is the corresponding vector of maximum likelihood estimates, obtained by fitting the incidence and survival models to the registry data. Such index is used as a correction factor of the LDP and yields the *Complete Prevalence* (CP) estimate:

$$\hat{N}_x(0, x; L) = \frac{\hat{N}_x(x - L, x)}{R_x(L; \hat{\psi})}. \qquad (3)$$

The CP therefore solves the bias due to the underestimation of the LDP, whenever the latter is observed.

2   The CHILDPREV method

In the case of childhood cancer there is a limited number of observations, only regarding the more recent years, and the LDP is zero for most of the adult ages. Therefore the CP cannot be computed by using (3). The method we propose is based on decomposing the cases diagnosed at age $[0, t_0]$ into the difference between cases diagnosed at age $[0, x]$ and those diagnosed at age $[t_0, x]$, and computing the corresponding prevalence by using the appropriate completeness index. The Lexis diagram in Figure 1, where each diagonal line represents the history of a patient through the age-and-year plane, provides an example. When current age $x > t_0 + L$ (i.e. patients were aged $x - L > t_0$ at the starting date of the observational period) no cases have been included in the registry; when $x \leq t_0 + L$ (i.e. patients were aged $x - L \leq t_0$ at the starting date of the registry) the observational period $[x - L, x]$ partially overlaps the period of interest $[0, t_0]$ and only a portion of cases are observed and already included in the registry. Those cases which were not counted are to be estimated.

Case 1: $x > t_0 + L$

The CP at current age x of cases diagnosed between ages 0 and t_0 is:

$$N_x(0, t_0) = N_x(0, x) - N_x(t_0, x). \qquad (4)$$
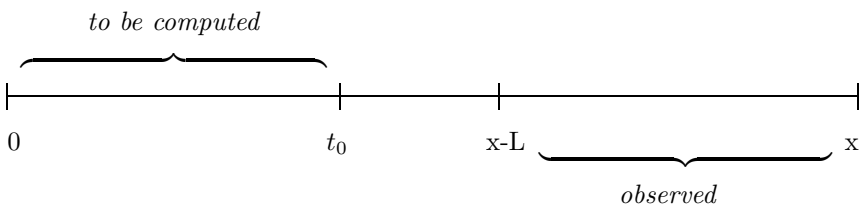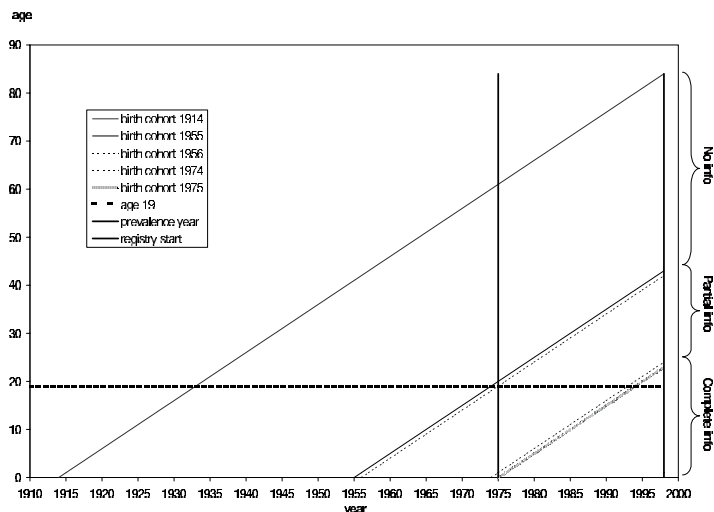
FIGURE 1. Lexis diagram; age upper limit $t_0 = 19$; data available from $1/1/1975$ to $1/1/1999$; prevalence computed on $1/1/1999$. Birth cohorts 1975–98 yield complete information; birth cohorts before 1956 no information; birth cohorts 1956–74 contribute only if they became ill at age 19 or less, after $1/1/1975$.

Here $N_x(0, x)$ is estimated by (3), while $N_x(t_0, x)$ is estimated as $\hat{N}_x(t_0, x; L) = \frac{\hat{N}_x(x-L, x)}{R_x^*(L; \hat{\psi})}$, the "complete" prevalence restricted to the age interval $[t_0, x]$, and the partial completeness index $R_x^*(L; \hat{\psi})$ is obtained as the ratio of two completeness indices $R_x^*(L; \hat{\psi}) = \frac{R_x(L; \hat{\psi})}{R_x(x-t_0; \hat{\psi})}$. Substituting $\hat{N}_x(0, x; L)$, $\hat{N}_x(t_0, x; L)$ and $R_x^*(L; \hat{\psi})$ in (4) we obtain

$$\hat{N}_x(0, t_0; L) = \frac{\hat{N}_x(x - L, x)}{R_x(L; \hat{\psi})} \left[1 - R_x(x - t_0; \hat{\psi}) \right]. \tag{5}$$

Case 2: $x \leq t_0 + L$

The period of interest $[0, t_0]$ and the observational period $[x-L, x]$ partially overlap, hence some cases are registered and some need to be estimated.
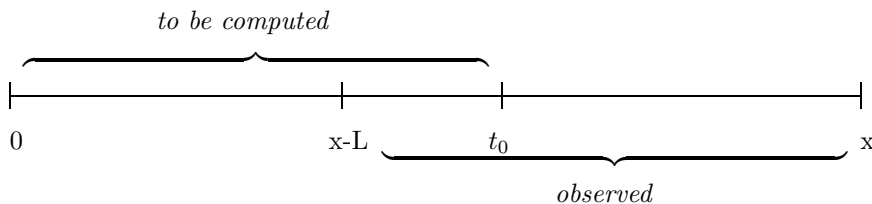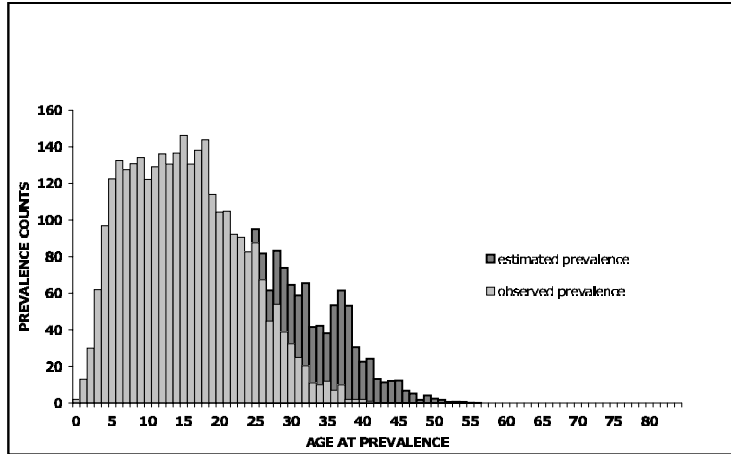
FIGURE 2. Estimated complete prevalence of Acute Lymphocytic Leukemia diagnosed in childhood age via the CHILDPREV method.

The prevalence of interest is

$$N_x(0, t_0) = N_x(0, x - L) + N_x(x - L, t_0), \qquad (6)$$

where the first (unobserved) summand is estimated as the difference between the complete and the observed prevalence, and the second (observed) summand is a fraction of the observed prevalence. Hence the estimated prevalence is

$$\hat{N}_x(0, t_0; L) = \frac{\hat{N}_x(x - L, x)}{R_x(L; \hat{\psi})} \left[1 - R_x(L; \hat{\psi})\right] + \sum_{t=x-L}^{t_0} N_x(t), \qquad (7)$$

where $N_x(t)$ are the observed prevalent cases of current age x who were diagnosed at age $t \in (x - L, \ldots, t_0)$.

3   An application

The CHILDPREV method has been applied to data collected by 9 US cancer registries (SEER9) to estimate the prevalence of adult patients who had been diagnosed with Acute Lymphocytic Leukemia (ALL) in the age interval [0,19]. Data for the period 1/1/1975 through 1/1/1999 were provided by the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute. Results are illustrated in Figure 2: the dark part

of the histogram denotes the estimated cases, the light part the observed ones. Up until the age of 24 the registries contain complete information on the patients; between age 25 and 43 cases are observed if they became ill after 1975, and are estimated if they became ill before 1975; between age 44 and 58 cases are completely estimated; after age 58 there are no cases at all, since children who became ill before 1960 did not survive (Mauer and Simone, 1976). With this method we estimate an extra 25% of cases which were not included in the LDP, but are still alive.

4   Discussion

The CHILDPREV method is based on the *Completeness Index*, which has been successfully implemented in the estimation of complete prevalence for various cancer sites in the US (Mariotto, *et al.*, 2002). It relies on modelling assumptions regarding the past behaviour of the disease. In the case of ALL we consider a survival model with cure (De Angelis *et al.*, 1999), and assume that only a portion of patients will die with a relative survival following a Weibull distribution, while the remaining have the same mortality rate as the general population; moreover, we assume that the survival function is zero for all cases diagnosed before 1960, regardless of their age. For the incidence function, which describes the relationship between cancer incidence and age, we adopt the model proposed by Merrill *et al.* (2000), which assumes a logistic function having as regressor a sixth degree polynomial function of age. The sensitivity of R to both models has been extensively studied by Capocaccia and De Angelis (1997).

References

Capocaccia, R. and De Angelis, R. (1997). Estimating the completeness of prevalence based on cancer registry data. *Statistics in Medicine*, **16**.

De Angelis, R., Capocaccia, R., Hakulinen, T., Soderman, B., Verdecchia, A. (1999). Mixture models for cancer survival analysis: application to population-based data with covariates. *Statistics in Medicine*, **18**.

Mariotto, A., Gigli, A., Capocaccia, R., Tavilla, A., et al. (2002). Complete and Limited Duration Cancer Prevalence Estimates. *SEER Cancer Statistics Review 1973-1999.* National Cancer Institute.

Mauer, A.M. and Simone, J.V. (1976). The current status of the treatment of childhood acute lymphoblastic leukemia. *Cancer Treatment Reviews*, **3**.

Merrill, R.M., Capocaccia, R., Feuer, E.J., Mariotto, A. (2000). Cancer prevalence estimates based on tumour registry data in the Surveillance, Epidemiology, and End Results (SEER) Program. *International Journal of Epidemiology*, **29**.

Chemical balance weighing designs with correlated errors based on balanced block designs

Bronisław Ceranka[1] and Małgorzata Graczyk[2]

[1] State School of Higher Vocational Education, Mickiewicza 5, 64-100 Leszno, Poland, e-mail: bronicer@owl.au.poznan.pl
[2] Department of Mathematical and Statistical Methods Agricultural University, Wojska Polskiego 28, 60-637 Poznań, Poland, e-mail: magra@owl.au.poznan.pl

Abstract: The paper is studying the estimation problem of individual measurements (weights) of objects using the chemical balance weighing design under the restriction on the number of times in which each object is weighed. We assume that the errors are correlated and they have equal variances. We give the lower bound of variance of each of the estimators and the sufficient and necessary conditions under which this lower bound is attained. The new construction method for the optimum chemical balance weighing design is given. We use the incidence matrices of the balanced incomplete block designs and the ternary balanced block designs to construct the design matrix of the optimum chemical balance weighing designs.

Keywords: balanced incomplete block design; chemical balance weighing design; ternary balanced block design.

1   Introduction

The results of n weighing operations aimed at determining the individual weights of p objects with a balance corrected for bias will fit into the linear model

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e},$$

where $\mathbf{y}$ is an $n \times 1$ random column vector of the observed weights, the design matrix $\mathbf{X}$ belongs to the class of $n \times p$ matrices of elements equal to $-1, 0$ or 1 and in which maximum number of elements equal to -1 and 1 in each column is equal to m, i.e. $\mathbf{X} \in \mathbf{\Phi}_{n \times p, m}(-1, 0, 1)$, $\mathbf{w}$ is an $p \times 1$ column vector representing unknown weights of objects and $\mathbf{e}$ is an $n \times 1$ random column vector of errors such that $E(\mathbf{e}) = \mathbf{0}_n$ and $E(\mathbf{e}\mathbf{e}') = \sigma^2\mathbf{G}$, where $\mathbf{0}_n$ is an $n \times 1$ column vector of zeros,

$$\mathbf{G} = g\left[(1-\rho)\mathbf{I}_n + \rho\mathbf{1}_n\mathbf{1}_n'\right], \quad g > 0, \quad \frac{-1}{n-1} < \rho < 1. \tag{1}$$

Now, if $\mathbf{X}'\mathbf{G}^{-1}\mathbf{X}$ is nonsingular, the least squares estimator of $\mathbf{w}$ is given by

$$\hat{\mathbf{w}} = (\mathbf{X}'\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{G}^{-1}\mathbf{y}$$

and the variance - covariance matrix of $\hat{\mathbf{w}}$ is of the form

$$\mathrm{Var}(\hat{\mathbf{w}}) = \sigma^2(\mathbf{X}'\mathbf{G}^{-1}\mathbf{X})^{-1}.$$

In the case $\mathbf{G} = \mathbf{I}_n$, some problems connected with optimum chemical balance weighing designs have been studied in Hotelling (1944), Raghavarao (1971), and Banerjee (1975). In the situation when not all objects are included in each weighing operation and the errors are correlated with equal variances, the problem of existing of the optimum chemical balance weighing design was considered in Ceranka and Graczyk (2003). They have given the lower bound of variance of each of the estimators and the definition of the optimal design. In the same paper they have given the necessary and sufficient conditions under which the chemical balance weighing design with the design matrix $\mathbf{X} \in \mathbf{\Phi}_{n \times p, m}(-1, 0, 1)$ and with the variance-covariance matrix of errors $\sigma^2\mathbf{G}$, where $\mathbf{G}$ is of the form (1) is optimal. Hence, from Ceranka and Graczyk (2003) we have

Theorem 1. Let $0 \leq \rho < 1$. Any nonsingular chemical balance weighing design with the design matrix $\mathbf{X} \in \mathbf{\Phi}_{n \times p, m}(-1, 0, 1)$ and with the variance-covariance matrix of errors $\sigma^2\mathbf{G}$, where $\mathbf{G}$ is given in (1), is optimal if and only if

$$\mathbf{X}'\mathbf{X} = m\mathbf{I}_p \qquad \text{and} \qquad \mathbf{X}'\mathbf{1}_n = \mathbf{0}_p. \tag{2}$$

Theorem 2. Let $\frac{-1}{n-1} < \rho < 0$. Any nonsingular chemical balance weighing design with the design matrix $\mathbf{X} \in \mathbf{\Phi}_{n \times p, m}(-1, 0, 1)$ and with the variance-covariance matrix of errors $\sigma^2\mathbf{G}$, where $\mathbf{G}$ is given in (1), is optimal if and only if

$$\mathbf{X}'\mathbf{X} = m\mathbf{I}_p - \frac{\rho(m - 2u)^2}{1 + \rho(n - 1)}(\mathbf{I}_p - \mathbf{1}_p\mathbf{1}'_p),$$

$$u_1 = u_2 = ... = u_p = u, \tag{3}$$

and
$$\mathbf{X}'\mathbf{1}_n = \mathbf{z}_p,$$

where $u = min(u_1, u_2, ..., u_p)$, u_j represents the number of elements equal to -1 in the jth column of the matrix $\mathbf{X}$, $\mathbf{z}_p$ is $p \times 1$ vector, for which jth element is equal to $(m - 2u)$ or $-(m - 2u)$, $j = 1, 2, ..., p$.

But, in Ceranka and Graczyk (2003) were some methods of construction of the design matrix $\mathbf{X} \in \mathbf{\Phi}_{n \times p, m}(-1, 0, 1)$ not given. Because of this reason in present paper we give the method of construction of the design matrix $\mathbf{X} \in \mathbf{\Phi}_{n \times p, m}(-1, 0, 1)$. It is based on the incidence matrices of the balanced incomplete block designs and the ternary balanced block designs.

2   Construction of the design matrix

Let $\mathbf{X} \in \mathbf{\Phi}_{n \times p,m}(-1, 0, 1)$ be the design matrix of the chemical balance weighing design given in the form

$$\mathbf{X} = \left[\begin{array}{c} 2\mathbf{N}_1' - \mathbf{1}_{b_1}\mathbf{1}_v' \\ \mathbf{N}_2' - \mathbf{1}_{b_2}\mathbf{1}_{v'} \end{array} \right], \tag{4}$$

where $\mathbf{N}_1$ is the incidence matrix of the balanced incomplete block design with the parameters v, b_1, r_1, k_1, λ_1 (see Raghavarao (1971)) and $\mathbf{N}_2$ is the incidence matrix of the ternary balanced block design with the parameters v, b_2, r_2, k_2, λ_2, ρ_{12}, ρ_{22} (see Billington (1984)).

Lemma 1. The chemical balance weighing design with the matrix $\mathbf{X} \in \mathbf{\Phi}_{n \times p,m}(-1, 0, 1)$ given in the form (4) is nonsingular if and only if

$$2k_1 \neq k_2 \quad \text{or} \quad 2k_1 = k_2 \neq v.$$

The optimality conditions given in Ceranka and Graczyk (2003) are depended on the parameter ρ which is connected with the matrix $\mathbf{G}$. This implies   that the methods of   construction   of the design matrix $\mathbf{X} \in \mathbf{\Phi}_{n \times p,m}(-1, 0, 1)$ are depended on ρ, either. Hence we have

Theorem 3. Let $0 \leq \rho < 1$. Any nonsingular chemical balance weighing design with the matrix $\mathbf{X} \in \mathbf{\Phi}_{n \times p,m}(-1, 0, 1)$ given by (4) and with the variance - covariance matrix of errors $\sigma^2 \mathbf{G}$, where $\mathbf{G}$ is of the form (1), is optimal for estimation unknown measurements of objects if and only if

$$b_1 - 4(r_1 - \lambda_1) + b_2 + \lambda_2 - 2r_2 = 0. \tag{5}$$

and

$$b_1 - 2r_1 + b_2 - r_2 = 0. \tag{6}$$

Proof. For the design matrix $\mathbf{X} \in \mathbf{\Phi}_{n \times p,m}(-1, 0, 1)$ in the form (4) we have

$$\mathbf{X}'\mathbf{X} = [4(r_1 - \lambda_1) + r_2 + 2\rho_{22} - \lambda_2]\mathbf{I}_v + \eta\mathbf{1}_v\mathbf{1}_v', \tag{7}$$

where $\eta = b_1 - 4(r_1 - \lambda_1) + b_2 + \lambda_2 - 2r_2$. Then for $0 \leq \rho < 1$ from (7) and (2) it derivers that the conditions (5) and (6) are true.

Theorem 4. Let $\frac{-1}{n-1} < \rho < 0$. Any nonsingular chemical balance weighing design with the matrix $\mathbf{X} \in \mathbf{\Phi}_{n \times p,m}(-1, 0, 1)$ given by (4) and with the variance - covariance matrix of errors $\sigma^2 \mathbf{G}$, where $\mathbf{G}$ is of the form (1), is optimal if and only if

$$\rho = \frac{\eta}{(2r_1 - b_1 + r_2 - b_2)^2 - \eta(b_1 + b_2 - 1)} \tag{8}$$

and

$$\eta < 0. \tag{9}$$

Proof. From the theorem (2) it derivers that the chemical balance weighing design $\mathbf{X} \in \mathbf{\Phi}_{n \times p, m}(-1, 0, 1)$ in the form (4) with the variance-covariance matrix of errors $\sigma^2 \mathbf{G}$, where $\mathbf{G}$ is of the form (1), is optimal if and only if the conditions (3) are true. From the last one of them it follows that $\mathbf{z}_p$ is equal to $(m - 2u)$ or $-(m - 2u)$, where $m - 2u = 2r_1 - b_1 + r_2 - b_2$. Now from the first condition of (3) and from (7) we have $\eta = \frac{\rho(2r_1 - b_1 + r_2 - b_2)^2}{1 + \rho(b_1 + b_2 - 1)}$, which complete the proof.

3   The Examples

3.1   Example 1

Let us consider the estimation problem of $p = 16$ objects using $n = 48$ measurement operations. We assume that each object is weighed at least $m = 24$ times. The variance - covariance matrix of errors $\sigma^2 \mathbf{G}$ is given by the matrix $\mathbf{G}$ of the form (1) with $0 \leq \rho < 1$. For estimation of unknown measurements of objects we use the optimum chemical balance weghing design with the design matrix $\mathbf{X} \in \mathbf{\Phi}_{48 \times 16, 24}(-1, 0, 1)$ given by the formula (4). To construct the design matrix we use the incidence matrix of the balanced incomplete block design with the parameters $v = 16$, $b_1 = 16$, $r_1 = 10$, $k_1 = 10$, $\lambda_1 = 6$ given through blocks (4,5,6,7,8,9,10,11,14,15), (3,4,5,7,8,11,12,13,14, 16), (2,4,5,9,10,11,12,13,15, 16), (2,3,6,8,9,10,11,12,13,16), (2,3,5,6,7,9,12,14, 15,16), (2,3,4,6,8,10,13,14, 15,16), (1,7,8,9,10,12,13,14,15,16), (1,3,5,6,7,10,11, 13,15,16), (1,3,4,6,8,9, 11,12,15,16), (1,3,4,5,6,9,10,12,13,14), (1,2,5,6,7,8,9, 11,13,14), (1,2,4,6,7,10, 11,12,14,16), (1,2,4,5,6,7,8,12,13,15), (1,2,3,5,8,10, 11,12,14,15), (1,2,3,4,7,9, 11,13,14,15), (1,2,3,4,5,7,8,9,10,16) and the incidence matrix of the ternary balanced block design with the parameters $v = 16$, $b_2 = 32$, $r_2 = 28$, $k_2 = 14$, $\lambda_2 = 24$, $\rho_{12} = 24$, $\rho_{22} = 2$ $\mathbf{N}_2 = [\mathbf{A} \vdots \mathbf{A}]$, where $\mathbf{A} = \mathbf{1}_{16} \mathbf{1}_{16}' + [\mathbf{I}_4 \otimes (2\mathbf{I}_4 - \mathbf{1}_4 \mathbf{1}_4')]$, where $\otimes$ denotes the Kronecker product of the matrices. Thus, the design matrix $\mathbf{X} \in \mathbf{\Phi}_{48 \times 16, 24}(-1, 0, 1)$ is optimal and permits for estimation of unknown measurements of objects with minimal variance equal to $Var(\hat{w}_j) = \frac{\sigma^2 g(1-\rho)}{24}$ for each $0 \leq \rho < 1$ and $g > 0$, $j = 1, 2, .., 16$.

3.2   Example 2

For $\frac{-1}{n-1} < \rho < 0$ we consider the estimation problem of $p = 5$ objects using $n = 15$ measurement operations. We assume that each object is weighed at least $m = 14$ times. The variance - covariance matrix of errors $\sigma^2 \mathbf{G}$ is given by the matrix $\mathbf{G}$ of the form (1) with $\rho = -\frac{3}{46}$. For estimation of unknown measurements of objects we use the optimum chemical balance weghing design with the design matrix $\mathbf{X} \in \mathbf{\Phi}_{15 \times 5, 14}(-1, 0, 1)$ given by the formula (4). To construct the design matrix $\mathbf{X} \in \mathbf{\Phi}_{15 \times 5, 14}(-1, 0, 1)$ of the optimum chemical balance weighing design in the form (4) we use

the incidence matrix $\mathbf{N}_1$ of the balanced incomplete block design with the parameters $v = 5$, $b_1 = 10$, $r_1 = 4$, $k_1 = 2$, $\lambda_1 = 1$ and the incidence matrix $\mathbf{N}_2$ of the ternary balanced block design with the parameters $v = 5$, $b_2 = 5$, $r_2 = 5$, $k_2 = 5$, $\lambda_2 = 4$, $\rho_{12} = 1$, $\rho_{22} = 2$, where

$$\mathbf{N}_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}, \; \mathbf{N}_2 = \begin{bmatrix} 1 & 2 & 2 & 0 & 0 \\ 2 & 1 & 0 & 2 & 0 \\ 2 & 0 & 1 & 0 & 2 \\ 0 & 2 & 0 & 1 & 2 \\ 0 & 0 & 2 & 2 & 1 \end{bmatrix}.$$

We can show that the optimality condition (3) is given by $\mathbf{X}'\mathbf{X} = 17\,\mathbf{I}_5$ $-3\mathbf{1}_5\mathbf{1}_5'$. Thus, the design matrix $\mathbf{X} \in \mathbf{\Phi}_{15 \times 5, 14}(-1, 0, 1)$ is optimal and permits for estimation of unknown measurements of objects with minimal variance equal to $Var(\hat{w}_j) = \frac{49\sigma^2 g}{46 \cdot 17}$ for each $g > 0$, $j = 1, 2, 3, 4, 5$.

References

Banerjee, K.S. (1975). *Weighing Designs for Chemistry, Medicine, Economics, Operations Research, Statistics.* Marcel Dekker Inc.,New York.

Billington, E. J. (1984). *Balanced n-array designs: a combinatorial survey and some new results* Ars Combin. 17A, 37-72.

Ceranka B. and Graczyk M. (2003). *On the estimation of parameters in the chemical balance weighing designs under the covariance matrix of errors $\sigma^2\mathbf{G}$.* 18th Interational Workshop on Statistical Modelling, 69-74.

Hotelling, H. (1944). *Some improvements in weighing and other experimental techniques.* Ann. Math. Stat., 15, 297-305.

Raghavarao, D. (1971). *Constructions and Combinatorial Problems in Design of Experiments.* New York: John Wiley Inc.

Rao, C.R. (1973). *Linear Statistical Inference and its Applications.* New York: John Wiley and Sons Inc.

The forward search for generalised extreme value distributions

Fabrizio Laurini, Aldo Corbellini

[1] Department of Economics – Division of Statistics and Computing, University of Parma, Via J.F. Kennedy 6, 43100 Parma, Italy.

Abstract: Statistical model selection, based on the likelihood ratio test, can be biased due to the presence of few influential observations or some model mis-specification. A forward analysis of the data can help understanding model fitting failures and it gives new insights for model selection. In this paper we consider the model selection problem for distributions studied in extreme value analysis.

Keywords: extreme value models; likelihood ratio test; forward search.

1 Introduction

Nowadays, the statistical analysis of extreme values is of great concern in several fields as, for instance, hydrology, geology and finance. For predicting "what appears to be unpredictable", many researchers have switched their attention to develop some methods able to model common features shown by rare events. Many techniques are currently available for addressing such an issue, and they are collectively called *models for extreme values.*
In this paper we suppose to deal with data of so called *block maxima*, like, for instance, the maxima of a monthly return of some asset price. The class of generalised extreme value (GEV) distributions is suited to model block maxima, and has distribution function

$$G(x) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\}, \quad \text{for } \{x : 1 + \xi(x - \mu) > 0\} \quad (1)$$

where $\{x\}_+ = \max(0, x)$, $\sigma > 0$ and μ, σ, ξ are location, scale and shape parameters respectively; for details see Coles (2001). From expression (1), Fréchet and Weibull distributions arise for $\xi > 0$ and $\xi < 0$ respectively. The subset of the GEV family with $\xi = 0$, which is formally the limit $\xi \to 0$ of expression (1), leads to Gumbel distribution with representation

$$G(x) = \exp\left\{-\exp\left(\frac{x-\mu}{\sigma}\right)\right\}, \quad \text{for } -\infty < x < \infty. \quad (2)$$

Through the inference on the shape parameter ξ it is possible to select alternative models for block maxima. Suppose that $\mathcal{M}_1$ is the GEV model with parameters ξ, σ and μ, while $\mathcal{M}_0$ is the Gumbel model, i.e. the GEV model with the constraint $\xi = 0$. Defining with $L_{\mathcal{M}_1}(\xi, \sigma, \mu)$ and $L_{\mathcal{M}_0}(\sigma, \mu)$ the likelihood of models (1) and (2), we analyse the likelihood ratio (LR)

$$\Lambda = \frac{\sup L_{\mathcal{M}_0}(\sigma, \mu)}{\sup L_{\mathcal{M}_1}(\xi, \sigma, \mu)}.$$

The statistic $-2\log(\Lambda)$ is distributed as a chi-square χ_1^2 with 1 degree of freedom. Such statistic is often adopted for model selection purposes.

Assuming independence of block maxima, the likelihood for GEV models can be easily derived (see Coles, 2001) and, though there are not analytical solutions, maximum likelihood (ML) estimates can be obtained by standard numerical optimization algorithms. The likelihood function for models $\mathcal{M}_1$ and $\mathcal{M}_0$ is far to be elliptical, and by profiling the likelihood it is achieved a good level of accuracy.

Although extremes cannot be called outliers, the fit of a GEV distribution to data (i.e. the estimate of $\hat{\xi}$, $\hat{\sigma}$ and $\hat{\mu}$) is sensitive to model mis-specification and impact of influential observations.

2   Forward algorithm for GEV models

To study the sensitivity of parameter estimates to model mis-specification, we simulate data from a GEV density and we adopt the forward analysis technique of Atkinson & Riani (2000). The forward algorithm explores the agreement of data with a specified null model. By an exhaustive search, the null model is initially fitted on an outlier-robust subsample. Proceeding the search, only units closer to the specified null model join the initial subsample. Thus, observations are added to the initial subset according to their agreement to the specified null model. Such an agreement is monitored through diagnostics during the forward search.

Atkinson & Riani (2000) give a forward algorithm for linear models. The inclusion of observations to the initial subset is based on the ordered model residuals. Our context is slightly different, and we need to update the forward algorithm as follows:

Initial subset. For a N-size sample we fit a GEV model to all the $\binom{N}{k}$ subsamples of size k. The fitting is carried through ML, and we found numerical problems for $k \leq 4$. Denote with $\hat{f}_j^s(x)$ the likelihood contribution given by the s-th unit (with $s = 1, \ldots, N$) when the j-th subsample is considered, with $j = 1, \ldots, \binom{N}{k}$. Thus, $\hat{f}_j^s(x)$ is the estimated density for the s-th observation which arise when $\hat{\xi}$, $\hat{\sigma}$ and $\hat{\mu}$ are estimated using the j-th subsample. Suppose to order the contribution to the likelihood function of each observation, i.e. consider the ordered estimated densities $\hat{f}_j^{(1)}(x) \leq \ldots \leq \hat{f}_j^{(s)}(x) \leq \ldots \leq \hat{f}_j^{(N)}(x)$. For any subset $j = 1, \ldots, \binom{N}{k}$ we
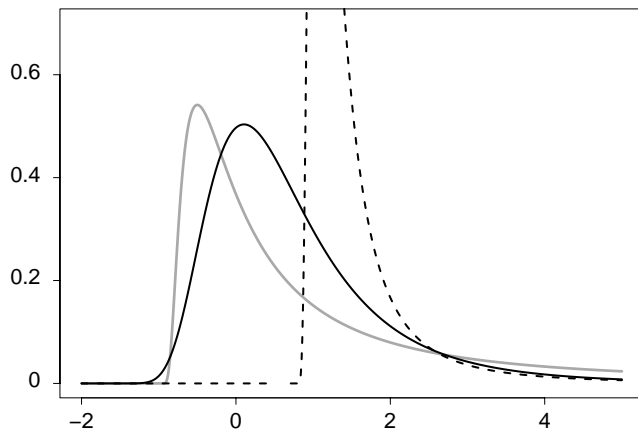
FIGURE 1. GEV densities. Gray line is the density of the model we generate from. Black dashed line is the density estimated by using data in $S^\star$. Black solid line is the density estimated using data at step 23 of the forward search.

sort such densities $\hat{f}_j^{(s)}(x)$. Denoting with "med" the sample median, we select the subsample $S^\star$ of size k which satisfies

$$\hat{f}_{S^\star}^{(\text{med})}(x) = \min_j \big(\hat{f}_j^{(\text{med})}(x)\big).$$

$S^\star$ should not be affected by the presence of influential observations.

Adding units. From step k to $k+1$ the unit joining $S^\star$ is such that its contribution to the likelihood to the fitted model is higher. At step $k+1$ a new model is fitted and new estimates $\hat{\xi}$, $\hat{\sigma}$ and $\hat{\mu}$ are obtained using the updated subsample of size $k+1$. This procedure is repeated until all units join the initial subset.

Monitoring statistics. During the forward search we monitor the behaviour of: i) LR test; ii) $\hat{\xi}$, $\hat{\sigma}$ and $\hat{\mu}$; iii) changes in the density of the null GEV model for any unit.

3   Example on simulated GEV data

We simulate 25 observations from a GEV distribution with parameters $\xi = 1$, $\mu = 0$ and $\sigma = 1$, and the density is sketched in Figure 1 (gray line). We select $S^\star$ by analysing all the $\binom{25}{5} = 53130$ subsamples. The estimated density based on data in subsample $S^\star$ is the dashed black line in Figure 1.
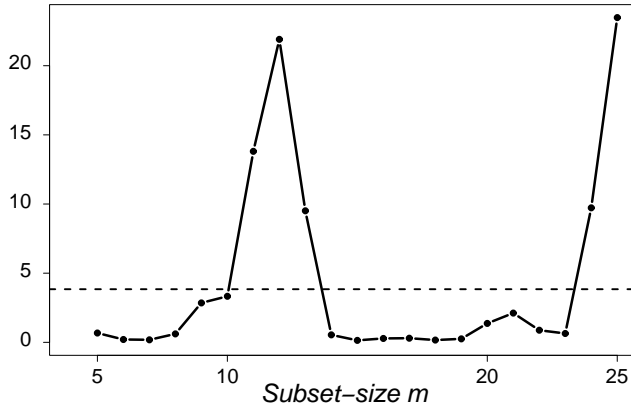
FIGURE 2. Behaviour of LR test $-2\log(\Lambda)$ during the forward search. Black dashed line is the 95-th quantile of a chi-square χ_1^2 with 1 degree of freedom.

In both cases the support of the distribution is bounded, with constraint induced by equation (1).

At each step of the forward search we monitor in Figure 2 the LR test $-2\log(\Lambda)$. When the value of such a test is smaller than a specified high quantile of a χ_1^2 distribution we should consider the model $\mathcal{M}_0$, as the inclusion of an extra parameter in the model (ξ, in our example) would not give enough contribution to the overall likelihood of model $\mathcal{M}_1$. For example, by inspecting Figure 2 at step 23, we would not accept the GEV model from which we truly generated the data. The density of the fitted model using the subsample at step 23 is the solid black line in Figure 1.

Finally, we also monitor the behaviour of ML estimate of ξ in Figure 3. At step 23 of the forward search we have $\hat{\xi} \approx 0.21$. At this step of the forward search we could not reject the hypothesis of dealing with the model $\mathcal{M}_0$, i.e. a Gumbel distribution with unbounded support.

4   Discussion

In this paper we provide an algorithm for the forward analysis of extreme value distributions, that provides new insights on the structure of GEV modeling. The main contribution consisted in updating the algorithm previously available for linear models. Decision are often made when the whole set of observations is available and, in practice, we showed that such decision can be highly sensitive to the presence of few influential observations.
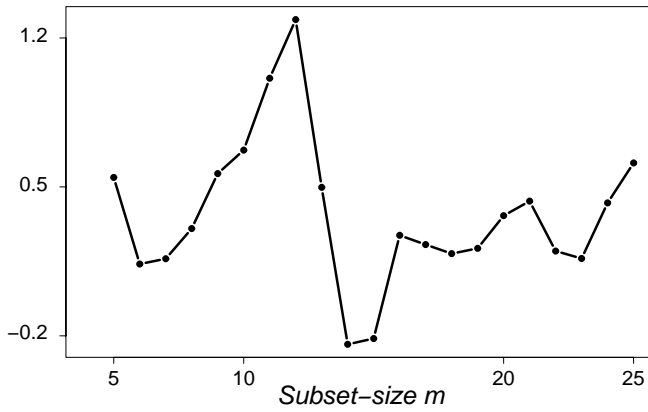
FIGURE 3. Behaviour of ML estimate of ξ during the forward search.

Future research could be oriented to study the behaviour of confidence intervals for diagnostic monitoring, updated at each step of the forward search.

References

Atkinson, A. C., Riani, M. (2000). *Robust Diagnostic Regression Analysis.* New York: Springer-Verlag.

Coles, S. G. (2001) *An Introduction to Statistical Modeling of Estreme Values.* London: Springer-Verlag.

Modelling Breast Cancer Data with Informative Dropout

Oskrochi, G.[1] and Crouchley, R.[2]

[1] Department of Mathematical Sciences, School of Technology, Oxford Brookes University. roskrochi@brookes.ac.uk
[2] Center for Applied Statistics, Lancaster University. r.crouchley@lancaster.ac.uk

Abstract: In building probabilistic models for survival times it is not always realistic to believe that all relevant risk factors or covariates are measured and included. Unmeasured or omitted risk factors often generate a between case variation usually referred to as frailty (in the biomedical literature), extra variation (in the statistical literature), and residual heterogeneity (in the social sciences literature). In order to properly interpret results of mutivariate survival analysis, one has to consider the fact that due to these frailties the individual risks may differ in unknown ways.

Keywords: Mixture models, Informative dropout, Multivariate frailty models.

1 Introduction

In frailty competing risk models, the observed changes in population hazard rates over time are a mixed result of two stochastic process: first, the actual changes in the individual hazards (i.e. observed risk factors), and, second, unobserved heterogeneity which causes the high-risk individual to have a shorter survival time. To understand the individual-level process, it is necessary to separate out these two effects. Moreover, the observed, population averaged, survival curves and hazard rates are difficult to interpret and potentially misleading.

These issues have been discussed by a number of authors, including, Lancaster and Nickell (1980), Stallard and Vaupel (1981), Heckman and Singer (1984, 1985), Vaupel and Yashin (1985), Hougaard (1984, 1986a, b), Aalen (1988, 1992) and Vaupel (1990). For illustration we consider the multiplicative frailty effects model which is commonly used in the literature. Let $f_{T|\tau}(t; \lambda|\tau)$ be the conditional density function of response time T at t with unknown parameter vector λ, given the unobserved frailty τ, such that λ is related to the conditional hazard by $h(t|\tau, \mathbf{X}) = \tau g(\lambda, \mathbf{X})$, where $\mathbf{X}$ is the observed covariate matrix and $\tau \sim P(\tau; \theta)$ with parameter vector θ. Unconditionally the marginal hazard has to be extracted from unconditional distribution of the response i.e. $f_T(t; \lambda) = \int f_{T|\tau}(t; \lambda|\tau) P(\tau; \theta) d\tau$. Many authors choose $P(\tau; \theta)$ as the conjugate of $f_{T|\tau}(t; \lambda|\tau)$ in order to

get a tractable form for $f_T(t; \lambda)$. If the marginal distribution is not analytically tractable, numerical integration or Monte Carlo simulations may be used. Alternatively the integration may be approximated by analytically tractable forms. There is no guarantee that the use of conjugate distribution for unobserved frailty τ is the best choice. One may use the central limit theorem to justify the use of a normal distribution as the distribution of the unobserved frailty when there is no prior knowledge about the nature of the frailty distribution. In a multivariate case the normal distribution also allows for a general correlation structure between the frailties.

1.1   A competing risk model for breast Cancer Recurrence

We illustrate $f_T(t; \lambda)$ on some breast cancer (BC) data. Diagnosis of recurrent cancer is more devastating or psychologically difficult for a woman than her initial breast cancer diagnosis, therefore the event of interest is the first recurrence time of BC patients after initial treatment, with AGE, STAGE of the disease at first diagnosis and the SURGERY TYPE, HISTOLOGY, and the cohort of initial Surgery as potential covariates. Once recurrent breast cancer has been detected, physicians will order additional tests to determine to what extent the cancer has spread. In **Local recurrence** cancerous tumor cells remain in the original site, and over time, grow back; but a **regional recurrence** of breast cancer is more serious than local recurrence because it usually indicates that the cancer has spread past the breast and into the axillary (underarm) lymph nodes and beyond. In addition to these two observed recurrence times we also consider the situations where the recurrence time is not observed because the patient was free of symptoms at the end date of the study (independent right censoring) and patient left, for some reason, before the end date of the study (dropped out).

In the former case, it is generally assumed that the censoring mechanism is independent of the recurrence times. However, this assumption may not apply to the latter; for instance, patients with severe sickness tend to have shorter survival time and are more likely to die from other disease due to general weakness. On the other hand, patients with minor problems after treatment may have very long or no relapse duration so that they may decide not to come back. Ignoring this fact and employing the commonly used estimation procedures underestimates the parameters of interest. We distinguish the following:

- Those patients who were alive at date last seen, with no disease, no recurrence (right censored failure time).

- Those who experienced the first local recurrence, LR, (T_1) .

- Those who experienced the first regional recurrence, RR, (T_2).

- Those who died from breast cancer (dropped out due to breast cancer) before the first recurrence, DB, (T_3) .

- Those who died from other causes (dropped out with other disease) before the first recurrence, DO, (T_4).

We construct a multivariate frailty model for competing risks of breast cancer recurrence, including two recurrence types and the above categories of dropout (a four dimensional frailty distribution using a Cholesky decomposition method, for more detail see Oskrochi and Davies,1997a, and b), and illustrate the consequences of ignoring the recurrence type and the dropout mechanisms.

1.2   Model Specification and Informative dropout

A semi-parametric Cox's proportional hazard model marginal fit to each latent failure time support a Weibull model for times to recurrence (T_1 and T_2), time to death from breast cancer (T_3), and time to death from other causes (T_4). Therefore we assume the following hazard models for T_k, k=1,2,3,4,

$$h_k(t) = \alpha t^{\alpha-1} \exp(\beta_{0k}) \exp(\beta_k' \mathbf{X}_k + \tau_k), \tag{1}$$

where $\tau = (\tau_1, \tau_2 , \tau_3, \tau_4)$ represents the unobserved specific individual effects and/or unobserved or unmeasured covariates of each response. The ith likelihood of this multivariate frailty model is now given by

$$L_i = \int_{\tau_1} \int_{\tau_2} \int_{\tau_3} \int_{\tau_4} \prod_{k=1}^{4} \left[[h_{k0}(t)\Psi(\mathbf{X}_k)]^{d_{ki}} \right]$$

$$S(t|\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4) f(\tau_1, \tau_2, \tau_3, \tau_4) d\tau_1 d\tau_2 d\tau_3 d\tau_4. \tag{2}$$

Where $\Psi(\mathbf{X}_k) = \exp(\beta_k' \mathbf{X}_k + \tau_k)$. We separate out the constants for notational convenience. A test of $h_{10}(t)\Psi(\mathbf{X}_1) = h_{20}(t)\Psi(\mathbf{X}_2)$, i.e. ignoring the constants, will be a test of whether we can collapse local and regional failure types. Some researchers (M. Dos Santos, et. al.) have treated the dropout due to the breast cancer as a recurrence of the breast cancer, i.e. they have assumed $h_{k0}(t)\Psi(\mathbf{X}_k) = h_{30}(t)\Psi(\mathbf{X}_3)$, $k = 1, 2$, which can be tested by $h_{k0}(t)\Psi(\mathbf{X}_k) \neq h_{30}(t)\Psi(\mathbf{X}_3)$, $k = 1, 2$. Some previous research (M. Dos Santos, et. al.) has assumed only two post-treatment states, recurrence, ($T_1 + T_2 + T_3$ in our term) and right censored failure time (right censored failure time and T_4in our term). For details of this kind of test in another context, see Bradley, Crouchley and Oskrochi (2003).

TABLE 1. Parameter Estimations. S: Sig.- NS: Not Sig.- LS: Less Sig.

Factor	Indep.	Chole.	indep.	Chole.
	Est. LR	Est. LR	Est. RR	Est. RR
LN(α)	-0.165	0.501	-0.160	0.612
CONST	-16.569	-20.769	-4.169	-8.435
AGE	-0.019	-0.04	-0.034	-0.079
STAGE	NS	NS	S	LS
Surgery type	S	S	S	NS
Histology	S	LS	S	LS
SURIN90	NS	NS	-0.421	-1.112
	Est. DB	Est. DB	Est. DO	Est. DO
LN(α)	-0.111	0.354	0.011	-0.019
CONST	-6.648	-9.581	-16.095	-16.341
AGE	0.013	0.013	0.089	0.091
STAGE	S	S	1.323	1.393
Surgery type	S	S	S	S
Histology	S	S	NS	NS
SURIN90	0.391	0.454	-0.355	-0.307

1.3   The Data

The data used in this study cover more than 3200 women referred to the Christie Hospital, U.K., by their GPs with breast cancer between 1985 and 1995, and their subsequent monitoring to 2001. This is an observational data set, hence, no randomization or clinical trial were involved. Note that recurrence is defined as what is clinically known as recurrence of breast cancer (i.e. after remission). If individual has left the study before observing her first recurrence the observation is right censored at the date last seen.

1.4   The Results

The results show that dropout due to breast cancer cannot be treated as the times to recurrence. The dropout from other causes is marginally informative about failure times via its random effects, and the failure times can not be pooled into one failure time when controlling for different treatment at initial diagnosis.

A deviance difference of 113 for 10 df. was obtained for heterogenous model over the independent model. The test of $h_{10}(t)\Psi(X_1) = h_{20}(t)\Psi(X_2)$ is rejected, with a deviance of 3421.72 for 17 df, i.e. local and regional failure types are different. The tests of $h_{k0}(t)\Psi(X_k) = h_{30}(t)\Psi(X_3)$, $k = 1, 2$, are also rejected with deviances of $d_1 = 212.6$ and $d_2 = 221.7$, both with 17 df, i.e. the dropout mechanisms from breast cancer cannot be treated as

time to recurrence. A test to collapse both type of death is also rejected with $d = 260.22$ with 17 df. This implies that we cannot assume that post treatment behaviour has only the states of recurrence and right censoring. The variance-covariance matrix of the random effects is:

$$
\begin{pmatrix}
\sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\
\sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\
\sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} \\
\sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^4
\end{pmatrix}
=
\begin{pmatrix}
8.4 & -0.50 & -0.18 & 0.01 \\
-0.50 & 9.59 & -1.96 & -0.57 \\
-0.18 & -1.96 & 4.82 & 0.44 \\
0.01 & -0.57 & 0.44 & 0.07
\end{pmatrix}
$$

This shows the frailty (unobserved heterogeneity) of death from breast cancer is weakly (negatively) associated with time to local recurrence, but it is strongly (negatively) associated with, a more serious, regional recurrence. The frailty of death from other causes is not associated with time to local recurrence, but it is strongly (negatively) associated with regional recurrence, and strongly (positively) associated with death from breast cancer. The nature of unobserved heterogeneity in this study is likely to be the patients' level of frailty. More frail patients are expected to have shorter survival time to both types of death, hence a positive (σ_{34}). The less frail patients are expected to have longer recurrence time, hence a negative association $(\sigma_{23}$ and $\sigma_{24})$.

A further complication is that the test for informative dropout $\sigma_{13} = \sigma_{14} = \sigma_{23} = \sigma_{24} = \sigma_{34} = 0$ has a deviance of 18.2 for 5 df. This test suggests that dropout is informative. In other words, we cannot perform a joint analysis of (T_1) and (T_2) and ignore what is happening to (T_3) and (T_4).

References

Aalen, O. O. (1992). *Modelling heterogeneity in survival analysis by the compound Poisson distribution*. The Annals of Applied Probability, Vol. 2, No, 4, 951-972.

Bradley, Crouchley and Oskrochi (2003). *Social exclusion and Labour market transition*, Journal of Labour Economics, 10.

Hougaard, P. (1986b). *Survival model for heterogeneous populations derived from stable distributions*, Biometrika, $387-396$.

M. Dos Santos, D; R. B. Davies and B. Francis (1995). *Nonparametric hazard versus nonparametric frailty distribution in modelling recurrence of breast cancer*. Journal of Statistical Planning and Inference, 47, 111-127.

Oskrochi, G. and R. B. Davies (1997a). *An EM-type algorithm for multivariate mixture models*. Statistics and Computing 7, 145-151.

Oskrochi, G. and R. B. Davies (1997b). *Stayers in mixed Markov renewal models*. Computational Statistics & Data Analysis 25 453-464.

Local Influence and Residual Analysis in Heteroscedastic Symmetrical Linear Models

Francisco José A. Cysneiros[1]

[1] Departamento de Estatística - CCEN, Universidade Federal de Pernambuco, Recife - PE 50749-540 - Brazil, e-mail: cysneiros@de.ufpe.br

Abstract: This work extends some diagnostics procedures to heteroscedastic symmetrical linear models. This class of models includes all symmetric continuous distributions, such as normal, Student-t, generalized Student-t, exponential power and logistic, among others. We present an iterative process for the parameter estimation and we derive the appropriate matrices for assessing the local influence under perturbation schemes. An standardized residual is deduced and illustrative example is given. S-Plus codes are available in the address www.de.ufpe.br/∼cysneiros/elliptical/ heteroscedastic.html to implement the author's method.

Keywords: Symmetrical distributions; Local influence; Residuals; Heteroscedastic models; Robust models.

1 Heteroscedastic symmetrical linear models

The problem of modelling variances has been discussed by various authors, particularly in the econometric area. Under normal error, for instance, Cook and Weisberg (1983) present some graphical methods to detect heteroscedasticity. Smyth (1989) describes a method which allows modelling the dispersion parameter in some generalized linear models. Moving away from normal error, let ϵ_i, $i = 1, \ldots, n$, be independent random variables with density function of the form

$$f_{\epsilon_i}(\epsilon) = \frac{1}{\sqrt{\phi_i}} g\{\epsilon^2/\phi_i\}, \ \epsilon \in I\!\!R, \tag{1}$$

where $\phi_i > 0$ is the scale parameter, $g : I\!\!R \to [0, \infty]$ is such that $\int_0^\infty g(u)du < \infty$. We shall denote $\epsilon_i \sim \mathrm{S}(0, \phi_i)$. The function $g(\cdot)$ is called density generator (see, for example, Fang, Kotz and Ng, 1990). We consider the linear regression model

$$y_i = \mu_i + \sqrt{\phi_i}\epsilon_i, \tag{2}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$ are the observed response values, $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ has values of p explanatory variables, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and $\epsilon_i \sim \mathrm{S}(0, 1)$. We have, when they exist, that $\mathrm{E}(Y_i) = \mu_i$ and $\mathrm{Var}(Y_i) =$

$\xi\phi_i$, where $\xi > 0$ is a constant given by $\xi = -2\varphi'(0)$, while $\varphi'(0) = d\varphi(u)/du|_{u=0}$ with $\varphi(\cdot)$ being a function such that $\varsigma(t) = e^{it\mu}\varphi(t^2\phi)$, $t \in I\!R$, where $\varsigma(t) = \mathrm{E}(e^{ity})$ is the characteristic function. We call the model defined by (1)-(2) heteroscedastic symmetrical linear model.

We assume that the dispersion parameter ϕ_i is parameterized as $\phi_i = h(\tau_i)$, where $h(\cdot)$ is a known one-to-one continuously differentiable function and $\tau_i = \mathbf{z}_i^T\boldsymbol{\gamma}$, where $\mathbf{Z}_i = (z_{i1}, \ldots, z_{iq})^T$ has values of q explanatory variables and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^T$. The function $h(\cdot)$ is usually called dispersion link function and it must be a positive-value function. One possible choice for $h(\cdot)$ is $h(\tau) = \exp(\tau)$. The dispersion covariates $\mathbf{z}_i$'s are not necessarily the same location covariates $\mathbf{x}_i$'s. It can be shown that $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are globally orthogonal parameters and the Fisher information matrix $\mathbf{K}$ for $\boldsymbol{\theta}$ is block-diagonal, namely $\mathbf{K} = \mathrm{diag}\{\mathbf{K}_\beta, \mathbf{K}_\gamma\}$. The Fisher information matrices $\mathbf{K}_\beta$ and $\mathbf{K}_\gamma$ for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are given by $\mathbf{K}_\beta = \mathbf{X}^T\mathbf{W}_1\mathbf{X}$ and $\mathbf{K}_\gamma = \mathbf{Z}^T\mathbf{W}_2\mathbf{Z}$, where $\mathbf{W}_1 = \mathrm{diag}\{4d_g/\phi_i\}$ and $\mathbf{W}_2 = \mathrm{diag}\{\frac{(4f_g-1)h_i'^2}{4\phi_i^2}\}$, for $i = 1, \ldots, n$, where $\mathbf{X}$ is a $n \times p$ matrix with rows $\mathbf{x}_i^T$, $v_i = -2W_g(u_i)$, $u_i = (y_i - \mu_i)^2/\phi_i$, $W_g(u) = \frac{g'(u)}{g(u)}$, $g'(u) = \frac{\partial g(u)}{\partial u}$, $h_i' = \frac{\partial h(\tau_i)}{\partial \tau_i}$ and $\mathbf{Z}$ is a $n \times q$ matrix with rows $\mathbf{z}_i^T$. An iterative process to get the maximum likelihood estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ may be developed by using, for example, the scoring Fisher method, which leads to the following system of equations:

$$\mathbf{X}^T\mathbf{W}_1^{(k)}\mathbf{X}\boldsymbol{\beta}^{(k+1)} = \mathbf{X}^T\mathbf{W}_1^{(k)}\mathbf{z}_\beta^{(k)} \quad \text{and} \quad \mathbf{Z}^T\mathbf{W}_2^{(k)}\mathbf{Z}\boldsymbol{\gamma}^{(k+1)} = \mathbf{Z}^T\mathbf{W}_2^{(k)}\mathbf{z}_\gamma^{(k)},$$

where $\mathbf{z}_\beta$ and $\mathbf{z}_\gamma$ are $n \times 1$ vectors whose components take the forms

$$z_{\beta_i} = \mu_i + \frac{v_i}{4d_g}(y_i - \mu_i) \qquad \text{and} \qquad z_{\gamma_i} = \tau_i + \frac{2\phi_i}{(4f_g - 1)h_i'}(v_i u_i - 1),$$

$d_g = \mathrm{E}\{W_g^2(U^2)U^2\}$ and $f_g = \mathrm{E}\{W_g^2(U^2)U^4\}$ with $U \sim \mathrm{S}(0,1)$. For example, the Student-t distribution with ν degrees of freedom one has $d_g = (\nu + 1)/4(\nu + 3)$ and $f_g = 3(\nu + 1)/4(\nu + 3)$.

2   Local influence

The idea behind local influence is concerned with the study of the behaviour of some influence measure around the vector of no perturbation $\boldsymbol{\omega_0}$. For example, if the likelihood displacement $\mathrm{LD}(\omega) = 2\{\mathrm{L}(\hat{\boldsymbol{\theta}}) - \mathrm{L}(\hat{\boldsymbol{\theta}}_\omega)\}$ is used, where $\hat{\boldsymbol{\theta}}_\omega$ denotes the maximum likelihood estimate under the perturbed model, the suggestion of Cook (1986) is to investigate the normal curvature of the lifted line $\mathrm{LD}(\omega_0 + a\boldsymbol{\ell})$, where $a \in I\!R$, around $a = 0$ for an arbitrary direction $\boldsymbol{\ell}$, $||\boldsymbol{\ell}|| = 1$. He shows that the normal curvature may be expressed in the general form $\mathrm{C}_\ell(\boldsymbol{\theta}) = 2|\boldsymbol{\ell}^T\boldsymbol{\Delta}^T\ddot{\mathbf{L}}_{\theta\theta}^{-1}\boldsymbol{\Delta}\boldsymbol{\ell}|$, where $\boldsymbol{\Delta}$ is a $(p + q) \times s$ matrix with elements $\Delta_{ij} = \partial^2\mathrm{L}(\boldsymbol{\theta}|\boldsymbol{\omega})/\partial\theta_i\partial\omega_j$, $i = 1, \ldots, p + q$ and $j = 1, \ldots, s$, with all the quantities evaluated at $\hat{\boldsymbol{\theta}}$.

Lesaffre and Verbeke (1998) suggest evaluating the normal curvature at the direction of the ith observation, that consists in evaluating $C_\ell(\boldsymbol{\theta})$ at the $n \times 1$ vector $\boldsymbol{\ell}_i$ formed by zeros with one at the ith position. Paula et al. (2003) discuss some diagnostics procedures in homoscedastic symmetrical nonlinear regression models. Suppose the log-likelihood function for $\boldsymbol{\theta}$ expressed as $L(\boldsymbol{\theta}|\boldsymbol{\omega}) = \sum_{i=1}^n \omega_i \log\{g(u_i)/\sqrt{\phi_i}\}$, where $0 \le \omega_j \le 1$ is a case weights. Under this perturbation scheme the matrix $\boldsymbol{\Delta}^T$ takes the form $\boldsymbol{\Delta}^T = [\mathbf{D}(\mathbf{g})\mathbf{D}(\mathbf{e})\mathbf{X}, \mathbf{D}(\mathbf{m})\mathbf{Z}]^T$ where $\mathbf{D}(\mathbf{g}) = \mathrm{diag}\{g_1, \dots, g_n\}$, $g_i = \frac{v_i}{\phi_i}$, $\mathbf{D}(\mathbf{m}) = \mathrm{diag}\{m_1, \dots, m_n\}$, $m_i = \frac{h'_i}{2\phi_i}(v_i u_i - 1)$, $\mathbf{D}(\mathbf{e}) = \mathrm{diag}\{e_1, \dots, e_n\}$ and $e_i = y_i - \mu_i$.

3   Local influence on predictions

Let $\mathbf{q}$ a $p \times 1$ vector explanatory variables values, for which we do not have necessarily an observed response. Then, the prediction at $\mathbf{q}$ is $\hat{\mu}(\mathbf{q}) = \sum_{j=1}^p q_j \hat{\beta}_j$. Analogously, the point prediction at $\mathbf{q}$ based on the perturbed model becomes $\hat{\mu}(\mathbf{q}, \omega) = \sum_{j=1}^p q_j \hat{\beta}_{jw}$, where $\hat{\boldsymbol{\beta}}_\omega = (\hat{\beta}_{1\omega}, \dots, \hat{\beta}_{p\omega})^T$ denotes the maximum likelihood estimate from the perturbed model. Thomas and Cook (1990) have investigated the effect of small perturbations on predictions at some particular point $\mathbf{q}$ in continuous generalized linear models. The objective function $f(\mathbf{q}, \boldsymbol{\omega}) = \{\hat{\mu}(\mathbf{q}) - \hat{\mu}(\mathbf{q}, \boldsymbol{\omega})\}^2$ was chosen due to simplicity and invariance with respect to scale change. The normal curvature at the unit direction $\boldsymbol{\ell}$ takes, in this case, the form $C_{\boldsymbol{\ell}} = |\boldsymbol{\ell}^T \ddot{\mathbf{f}} \boldsymbol{\ell}|$, where $\ddot{\mathbf{f}} = \partial^2 f/\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}^T = -2\boldsymbol{\Delta}^T(\ddot{\mathbf{L}}_{\beta\beta}^{-1}\mathbf{q}\mathbf{q}^T\ddot{\mathbf{L}}_{\beta\beta}^{-1})\boldsymbol{\Delta}$, is evaluated at $\boldsymbol{\omega}_0$ and $\hat{\boldsymbol{\beta}}$. One has that $\boldsymbol{\ell}_{max}(\mathbf{q}) \propto \boldsymbol{\Delta}^T \ddot{\mathbf{L}}_{\beta\beta}^{-1} \mathbf{q}$.

Consider an additive perturbation on the ith response, namely $y_{i\omega} = y_i + \omega_i s_i$, where s_i may be an estimate of the standardized deviation of y_i and $\omega_i \in I\!R$. Then, the matrix $\boldsymbol{\Delta}$ equals $\mathbf{X}^T \mathbf{D}(\mathbf{a})\mathbf{D}(\mathbf{s})$, where $\mathbf{D}(\mathbf{s}) = \mathrm{diag}\{s_1, \dots, s_n\}$ and $\mathbf{D}(\mathbf{a}) = \mathrm{diag}\{a_1, \dots, a_n\}$ $a_i = \frac{1}{\phi_i}\{v_i - 4W'_g(u_i)u_i\}$.. The vector $\boldsymbol{\ell}_{max}(\mathbf{q})$ is constructed here by taking $\mathbf{q} = \mathbf{x}_i$, which corresponds to the $n \times 1$ vector $\boldsymbol{\ell}_{max}(\mathbf{x}_i) \propto \mathbf{D}(\mathbf{s})\mathbf{D}(\mathbf{a})\mathbf{X}\{\mathbf{X}^T\mathbf{D}(\mathbf{a})\mathbf{X}\}^{-1}\mathbf{x}_i$. A large value for $\boldsymbol{\ell}_{max_i}(\mathbf{x}_i)$ indicates that the ith observation should have substantial local influence on $\hat{y}_i$. Then, the suggestion is to take the index plot of the $n \times 1$ vector $(\boldsymbol{\ell}_{max_1}(\mathbf{x}_1), \dots, \boldsymbol{\ell}_{max_n}(\mathbf{x}_n))^T$ in order to identify those observations with high influence on its own fitted value.

4   Residuals

Because we have a symmetrical class of errors it is reasonable to think on the residual $r_i = y_i - \hat{y}_i$ to perform residual analysis. A standardized version for r_i may be attained by using the expansions up to order n^{-1} by Cox and Snell (1968). After some algebraic manipulations, we find that

$$E(\mathbf{r}) = \mathbf{0} \quad \text{and} \quad \mathrm{Var}(\mathbf{r}) = \xi\boldsymbol{\Phi}\{\mathbf{I}_n - (4d_g\xi)^{-1}\mathbf{H}\},$$

where $\mathbf{H} = \mathbf{\Phi}^{-1/2}\mathbf{X}(\mathbf{X}^T\mathbf{\Phi}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Phi}^{-1/2}$ and $\mathbf{\Phi} = \text{diag}\{\phi_1, \ldots, \phi_n\}$, $\mathbf{I}_n$ is the identity matrix of order n, Therefore, a standardized form for r_i is given by

$$\mathrm{t}_{r_i} = \frac{(y_i - \hat{y}_i)}{\sqrt{\hat{\phi}_i \xi \{1 - (4d_g\xi)^{-1}\hat{h}_{ii}\}}}.$$

Simulation studies omitted here indicate that t_{r_i} has mean approximately zero, variance exceeding one, negligible skewness and some kurtosis.

5   Application

To illustrate an application we shall consider the data set described in Montgomery et al. (2001, Table 3.2). The interest is on predicting the amount of time required by the router driver to service of vending machines in an outlet. The service activity includes stocking the machine with beverage products and minor maintenance or housekeeping. They fitted a homoscedastic linear regression model with intercept where the response variable was the delivery time, y (min), the covariates were the number of cases of producted stocked (x_1) and the distance walked by the route driver (x_2) in a sample of 25 observations. In their diagnostic analysis, points 9 and 22 appear with large effects on the parameter estimates (see Montgomery et al. 2001, pp. 210,213,215,216,217). We propose to fit heteroscedastic linear models under error distributions with heavier tails than the normal ones, namely

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \sqrt{\phi_i}\epsilon_i, \qquad i = 1, \ldots, 25 \qquad (3)$$

with $\phi_i = \exp\{\alpha + \gamma x_{i2}\}$ and $\epsilon_i \sim S(0,1)$ mutually independent errors.
We tried various error distributions but only two models seem to fit the data as well as or better than the normal model, the Student-t with 4 degrees of freedom and the logistic-II models. The generated envelopes for the three postulated models do not present any unusual features, (see Figure 1). Figure 1 also presents the index plots of C_i under normal, Student-t and logistic-II errors. Influential observations appear in Student-t model with smaller values than normal and logistic-II models.

References

Cook, R.D. (1986) Assessment of local influence (with discussion). *Journal of the Royal Statistical Society, Series B*, **48**, 133-169.
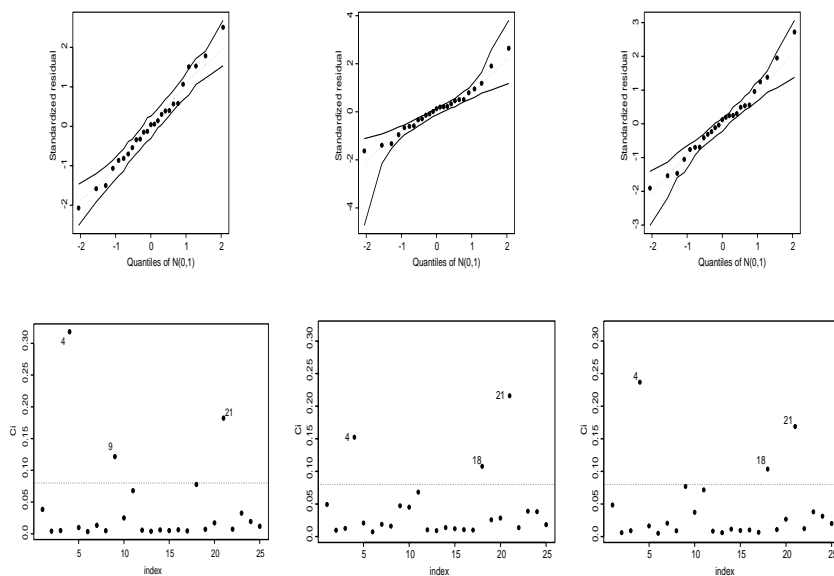
FIGURE 1. Envelopes and index plots of C_i under the normal (left), Student-t with 4 d.f. (middle) and logistic-II (right) fitted on the delivery data.

Cook, R.D. and Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika* **70**, 1-10

Cox, D.R. and Snell, E.J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society, Series B*, **30**, 248-275.

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

Lesaffre, F. and Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics* **54**, 579-582.

Montgomery, D.C., Peck, E.A. and Vining, G.G. (2001). *Introduction to Linear Regression Analysis*, 3rd ed. New York: Wiley.

Paula, G.A., Cysneiros, F.J.A. and Galea, M. (2003). Local influence and Leverage in elliptical Nonlinear Regression Models. In: *Proceedings of the 18th International Workshop on Statistical Modelling*; Verbeke, G., Molenberghs, G; Aerts, A. and Fieuws, S. (Eds). Leuven: Katholieke Universiteit Leuven, pp. 361-366.

Smyth, G.K. (1989). Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society, Series B*, **51**, 47-60.

Thomas,W. and Cook, R.D. (1990). Assessing influence on predictions from generalized linear models. *Technometrics* **32**, 59-65.

Modelling the costs of different strategies after myocardial infarction

Giulia Zigon[1], Alessandro Desideri[2] and Dario Gregori[3]

[1] Department of Statistics University of Florence
[2] Cardiovascular Research Foundation, Castelfranco Veneto
[3] Department of Public Health and Microbiology, University of Torino

Abstract: This study is aimed at evaluating the clinical factors and the management strategies, that affecting the hospitalization costs of the postinfarct patient. We use ordinary least square (OLS) linear regression, binary logistic regression, Cox proportional hazard model, parametric survival model assuming the Weibull distribution and the Aalen additive regression model. The mean predicted cost and the cost for specific clinical profile are compared. The Aalen model provides the most accurate prediction of mean cost and median cost (compared with the observed cost) and shows considerable promise for the analysis of the medical costs.

Keywords: Aalen additive regression model; Survival models; medical costs.

1 Introduction

Management of the postinfarct patient has changed in the last decade, aiming at the most cost-effective strategy; thus the study of the cost of the myocardial infarction (MI) and the factors affecting such cost are becoming more and more important for clinicians and policy-makers.
Risk stratification early after MI is an important goal in clinical decision making, because it allows to identify the high risk patients. In this connection different stratification modalities have been proposed: the simple clinical data obtained during the acute phase, the most commonly used exercise testing and more recently the coronary angiography and the stress echocardiography.
In the field of prognostic stratification it is still unclear what is the better choice between a invasive or not strategy in terms of cost-efficacy. Furthermore, the analysis of the medical costs presents several difficulties from the statistical point of view.
The data referring to the costs is characterized by a large mass of observations at zero cost, an asymmetric distribution, (because of a minority with high medical costs compared to the rest of the population) and the presence of dependent censoring (because of correlation between cost at censoring and cost-to-event) due to the patient deaths in the follow-up. The principal

methods used to analyze the effect of clinical factors on the medical costs (ordinary least square OLS, logistic regression) present problems connected to the inadequacy of the assumptions underlying the models.

According to the data characteristics and particularly to the presence of censoring, several works in literature (Dudley et al., 1993) have proposed to use the survival models like the Weibull model and the Cox regression model, because these models are based on few and/or more realistic assumptions concerning the distribution of the cost variable. Nevertheless the accrual of costs at different rates leads to dependent (or informative) censoring within subgroups defined by covariate levels and the proportional hazards (PH) assumption of these models is not in general satisfied (Etzioni et al., 1999).

The additive regression model (Aalen, 1989;1993) seems to be appealing, because it is not parametric (in the sense that functions, not parameters are fitted) and robust for the non proportional hazard and therefore an alternative to the Cox regression model.

On the basis of these considerations the purpose of this study is a comparison of analytic models for estimating the effect of clinical factors and management strategies on the costs of postinfarct patients. It is emphasized the innovative application of the Aalen additive regression model to medical costs and the performances of this model in terms of predicted costs.

2   Methods

2.1   The Data

A follow up of 1 year for medical costs was carried out in 10 General Hospital, eight in Italy and two in Turkey. Patients were admitted to the participating centers with a diagnosis of non complicated myocardial infarction, with beginning of the symptoms less than 24 hours, giving informed consent. For-hundred eighty-seven patients were enrolled and randomly assigned to three different strategies: 1) (132 patients) early use of pharmacological stress echocardiography under therapy (Day 3-5) and conventional discharge; 2) (130 patients) maximal symptom limited exercise testing under therapy, discharge in Day 7-9 ; 3) (225 patients) clinical evaluation and hospital discharge in Day 7-9. Cost of hospitalization was estimated referring to mean reimbursement for the diagnosis-related groups (DRG). Direct medical costs (in Euro) were calculated related to initial hospital stay, at 1, 6 months and 1 year follow-up. Total costs per patients were measured as the sum of initial hospital costs and follow-up hospital and outpatients costs (Figure 1). The clinical variables considered are age, gender, previous MI, diabetes, ejection fraction (EF), MI antero/lateral, strategy type.
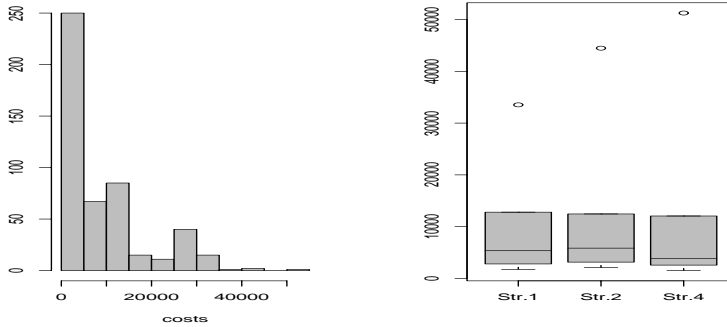
FIGURE 1. Cost distribution at 1 year of follw-up; b) Costs by strategies.

2.2   Models

Five different statistical models were applied:

- OLS linear regression

$$y = \sum a_i \ x_i \tag{1}$$

  where a_i are the regression coefficients and x_i the independent variables and y the observed costs.

- binary logistic regression

$$p(y > c) = \frac{1}{(1 + \ exp(- \sum a_i \ x_i))} \tag{2}$$

  where c is a fixed cutpoint (median and the third quartile) and $p(y > c)$ is the probability to have a cost greater than the median or the third quartile of the cost distribution.

- Parametric proportional hazard (PH) model assuming Weibull distribution. The Weibull p.d.f.

$$f(y) = \gamma\delta(y\delta)^{\gamma-1}exp[-(y\delta)^{\gamma}] \tag{3}$$

  where δ is the scale parameter and γ is the shape parameter can be extended to a regression model by allowing γ e δ to depend on $\mathbf{x}$, where $\mathbf{x}$ is a vector of covariates. The Weibull model can be written in the form

$$h(y|\mathbf{x}) = h_0(y)exp^{\mathbf{X}\beta} \tag{4}$$

  where $h(y|\mathbf{x})$ is the hazard function of the cost y given the covariates vector $\mathbf{x}$ and $h_0(y)$ is the baseline hazard function for the cost.

TABLE 1. Mean and Median of the predicted values by the models

	Obs. data	OLS m.	Weibull m.	Cox m.	Aalen m.
Mean	9162.152	9352.731	9938.465	9378	9281
Median	4845	9447.393	9822.701	4967	4556

- The Cox PH model. Considering the general form given in (4) in this model the regression coefficient is estimated in absence of knowledge of the baseline hazard function $h_0(y)$, that is the model is distribution free.

- The Aalen additive regression model

$$\lambda[y|Z] = \alpha_0 + \sum \alpha_k(y) \ Z_k \qquad (5)$$

where λ is the hazard rate of a cost y for an individual with a covariate vector Z_k; the hazard rate is a linear combination of the variables Z_k and $\alpha_k(y)$ are regression functions estimated from the data, which measure the influence of the respective covariates.

3   Results

Seven of the 487 patients died in the follow-up time, thus censoring is very low, about 1.4%. The normality assumption about the residuals for the OLS model is not satisfied (Shapiro-Wilk test p < 0.001). The cost data appears to obtain a good approximation with a Weibull distribution (scale parameter estimated=0.88), nevertheless the key assumption of proportional hazard of the Cox and Weibull models is not satisfied (Global ChiSquare: 22.88, p=0.005), particularly for age and strategy.
The considered clinical covariates are not significant except for the previous MI (yes) (Weibull model $p = 0.05$), the strategy 1 vs strategy 4 ($p = 0.01$ the logistic model with median cut-point) and the AMI location (antero-lateral)($p < 0.01$ for all the model except for the Aalen model $p = 0.05$). There is accord for all the considered models about this last variable, if we consider the third quartile (12319 euro) as cut-point of the logistic model. To compare the quantitative cost predictions of the models we computed the predicted costs relative to the mean and median (Table 1). The linear regression model OLS predicts enough well the mean cost, but overestimates the median and the same occurs for the Weibull model. The Cox model and the Aalen model perform well, particularly this last in the median value. The logistic regression predict well the proportion of costs greater than 12319 euro ($p = 0.26$) and 4845 euro ($p = 0.51$).
Finally, we compared the predicted costs for specific covariates values corresponding to different risk profiles from the clinical point of view (Table 2).

TABLE 2. Mean cost for specific clinical profile

Mean Observed cost	OLS model	Weibull model	Cox model	Aalen model
1) Male	Age>70	Previous AMI	EF<50%	strategy 2
5962.8	8269.8	7160.8	7157	6164
2) Female	Age< 70	MI antero-lateral	EF >50	strategy 4
9984	9516.4	9033	9193	9615

4   Discussion

The OLS model, the Weibull model and the Cox model underestimate slightly the medical cost for the second profile and overestimate the cost for the first profile (Table 2). The Aalen model (Aalen, 1989; 1993) is free from the PH assumption and performs better with respect to the other models, although there is an overestimation and an underestimation in the same direction as the others. The logistic model performs well (the extreme values do not influence the estimations) in predicting the high and low costs, but it precludes the computation of the mean cost and the choice of the dividing line (which is determinant in the analysis) is arbitrary. The Aalen additive regression model and the Cox model give a good estimation of the median with respect to the Weibull and the OLS models, that are sensitive to the high cost extreme values (Table 1). The accuracy of the Aalen model is superior to the accuracy of the other models in this dataset, but computer simulation studies will be necessary to establish the performance of this model in different circumstances.

References

Aalen, O.O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, **8**, 907-925.

Aalen, O.O. (1993). Further results on the non parametric linear regression model in survival analysis. *Statistics in Medicine*, **12**, 1569-1588.

Dudley, A.R., Harrell, F.E., Smith, R.L., Mark, D.B. et al. (1993). Comparison of anlytical models for estimating the effect of clinical factors on the cost of coronary artery bypass graft surgery. *Journal of Clinical Epidemiology*, **12**, 261-271.

Etzioni, R.D., Feuer, E.J., Sullivan, S., Lin, D., Hu, C. and Ramsey, S. (1999) On the use of survival analysis techniques to estimate medical care costs. *Journal of Health Economics*, **18**, 365-380.

Semiparametric Comparison of Two Samples

Konstantinos Fokianos

[1] University of Cyprus, Department of Mathematics & Statistics, P.O. Box 20537, Nicosia 1678, Cyprus

Abstract: We consider the density ratio model which specifies a linear parametric function of the log–likelihood ratio of two densities without assuming any specific form about them and has been found useful for semiparametric comparison of two samples. We study the Box–Cox family of transformations in the context of the density ratio model to suggest a data driven method for identification of the model's true parametric part. The methodology is illustrated by a real data example.

Keywords: biased sampling, semiparametrics, empirical likelihood

1 Introduction

Quite often in applications we come across with the problem of comparing two samples. The parametric theory resolves the question by appealing to the well known t-test. Accordingly, if $\{X_1, \ldots, X_{n_0}\}$ and $\{X_{n_0+1}, \ldots, X_n\}$ are two *independent* samples with $\bar{X}_0 = \sum_{i=1}^{n_0} X_i/n_0$ and $\bar{X}_1 = \sum_{i=n_0+1}^{n} X_i/n_1$ then it is well known that the two sample t-test rejects the hypothesis of means equality when

$$\frac{\bar{X}_0 - \bar{X}_1}{S\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}} \geq c \tag{1}$$

where

$$S^2 = \frac{\sum_{i=1}^{n_0} \left(X_i - \bar{X}_0\right)^2 + \sum_{i=n_0+1}^{n} \left(X_i - \bar{X}_1\right)^2}{n - 2},$$

and $n_1 = n - n_0$. The critical value c is determined by the t distribution with $n - 2$ degrees of freedom. To carry out test (1), both samples are assumed to be normally distributed with common unknown variance and unknown means.

Occasionally some (or all) of the needed assumptions fail so that (1) cannot be applied directly. A case in point is illustrated by Fig. 1(a) which displays boxplots of rainfall amounts from two groups of clouds. One group has been seeded with silver nitrate while the other has not. There is a total of 26 observations in each group and the purpose of the experiment was to
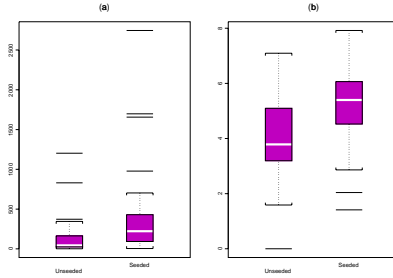
FIGURE 1. (a) Boxplots of the clouds data. (b) Boxplots of the clouds data after log transformation.

determine whether cloud seeding increases rainfall. The data are available at `http://lib.stat.cmu.edu/DASL/Stories/CloudSeeding.html`.
Figure 1(a) shows that both groups follow skewed distributions with large positive values. Clearly both assumptions of normality and equality of variances fail and therefore application of the two sample t-test is questionable. The problem may be bypassed after a logarithmic transformation which leads to symmetric distributions for both groups of clouds with approximately equal variances–see Fig. 1(b).
Here we consider a quite different approach to the two samples comparison problem. The methodology is relatively new and appeals on the so called *density ratio model* for *semiparametric* comparison of two samples. To be more specific assume that

$$
\begin{aligned}
X_1, ..., X_{n_0} &\quad \sim \quad f_0(x) \\
X_{n_0+1}, ..., X_n &\quad \sim \quad f_1(x) = \exp\left(\alpha + \beta h(x)\right) f_0(x).
\end{aligned}
\tag{2}
$$

where $f_i(x)$, $i = 0, 1$ are probability densities, h is a *known* function and α, β are two unknown parameters.
We refer to (2) as the density ratio model since it specifies a parametric function of the log likelihood ratio of two densities without assuming any specific form about them. Hence it is a semiparametric model and it is easy to see that under the hypothesis $\beta = 0$, both of the distributions are identical. Consequently if $\hat{\beta}$ stands for the maximum likelihood estimator of β (see (5)) then the following test procedure

$$
Z = \frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}}
\tag{3}
$$

where $\widehat{\mathrm{Var}(\hat{\beta})}$ denotes the estimated variance of $\hat{\beta}$, rejects the hypothesis $\beta = 0$ when $\mid Z \mid > c^{\star}$. The critical value $c^{\star}$ is determined by the

standard normal distribution. Recent contributions on semiparametric inference about the density ratio model include Qin and Zhang (1997), and more recently Fokianos et. all (2001).

2   Box–Cox Transformation for the Density Ratio Model

Recall (2) and assume that the data are positive, that is all $X > 0$. Assume that h is parameterized according the so called Box–Cox family of transformations

$$
h_\lambda(x) = \begin{cases} \dfrac{x^\lambda - 1}{\lambda} & \text{when } \lambda \neq 0 \\ \log x & \text{when } \lambda = 0. \end{cases}
$$

Thus expression (2) becomes

$$
\begin{aligned}
X_1, ..., X_{n_0} &\sim f_0(x) \\
X_{n_0+1}, ..., X_n &\sim f_1(x) = \exp\left(\alpha + \beta h_\lambda(x)\right) f_0(x).
\end{aligned} \tag{4}
$$

It turns out that the Box–Cox family of transformations enlarges the density ratio model by providing a data driven choice of $h(x)$. In this respect the data analyst can identify the appropriate $h(x)$ in applications. The following section discuss inference regarding model (4).

3   Inference

Inference can be carried out along the lines of Qin and Zhang (1997). Accordingly, it can be shown that inference for model (4) is based on the following empirical log likelihood

$$
l(\alpha, \beta, \lambda) = -\sum_{i=1}^n \log\left[1 + \rho_1 \exp\left(\alpha + \beta h_\lambda(x_i)\right)\right] + \sum_{i=n_0+1}^n \left(\alpha + \beta h_\lambda(x_i)\right), \tag{5}
$$

with $\rho_1 = n_1/n_0$. Expression (5) has been derived after profiling out an infinite dimensional parameter, namely the cumulative distribution function of $f_0(x)$, say $F_0(x)$. The key concept is that of the empirical likelihood (see Owen (1988)).

To estimate λ, maximize equation (5) for given λ with respect to α and β. If we denote by $l_{\max}(\lambda)$ the maximized log likelihood for a given value of λ, then a plot of $l_{\max}(\lambda)$ against λ for a trial series of values will reveal $\hat\lambda$–the maximum likelihood estimator of λ. An approximate $100(1-a)\%$ confidence interval for λ consists of those values of λ which satisfy the inequality

$$
l_{\max}(\hat\lambda) - l_{\max}(\lambda) \leq \frac{1}{2}\chi^2_{1;1-a} \tag{6}
$$
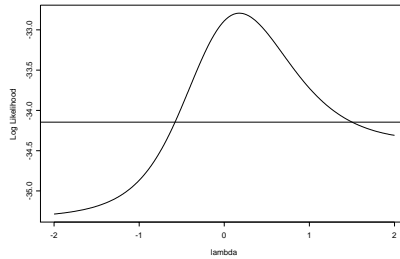
FIGURE 2. Values of the log likelihood for the clouds data when λ varies in $[-2, 2]$. The horizontal line indicates a 90% confidence interval for λ.

where $\chi^2_{1;1-a}$ is the percentage point of the chi–squared distribution with one degree of freedom.

3.1   Application

Figure 2 illustrates the above methodology applied to clouds data. In other words this is a plot of the maximized log likelihood as λ varies in $[-2, 2]$ with step equal to 0.01. The maximum value is obtained at $\hat{\lambda} = 0.18$. The horizontal line indicates a 90% confidence interval for λ–according to (6)– which turns out to be $[-0.58, 1.50]$. Consequently, values of λ equal to -1/2, 0, 1/2, 1 and 3/2 are not excluded as possibilities by the data. Apparently the relative small number of observations lead to negligible changes to the log likelihood for different λ and therefore the obtained confidence interval is rather large. Hence it is preferable to use values that fall near the viscosity of the maximum. For the clouds data we choose $\lambda = 0, 1/2$. This discussion confirms from another point of view that log transformation is appropriate for the data at hand.

References

Fokianos, K., Kedem, B., Qin, J., and Short, D. (2001). A semiparametric approach to the one–way layout. *Tecnometrics*, **43**, 56–65.

Qin, J., and Zhang, B. (1997). A goodness of fit test for logistic regression model based on case control data. *Biometrika*, **85**, 619–630.

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

Analysis of interval-censored data: A simulation study

Arminda Lucia Siqueira[1] and Inara Kellen Fonseca[1]

[1] Departamento de Estatística, ICEx, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil. E-mail: arminda@est.ufmg.br

Abstract: In many studies in which the response variable is the time until the occurrence of an event, the exact time cannot be determined, that is, only the interval of the occurrence is known. Such data can be analyzed by the traditional life table method (LTM) when there is no covariate. A more general approach consists in using a discrete-time regression model, such as proportional hazard model (DCM) or proportional odds model (DLM). In this paper we compare those three types of analysis (LTM, DCM, DLM) for the two-sample case through a simulation study. We assess the agreement among them with respect to the comparison between two groups, as well as the empirical power and the length of the confidence interval for quantities of interest. We also investigate the impact of the misspecification of the regression model.

1   Introduction

In several studies, the main outcome is the time between the beginning of the observation and the occurrence of an event of interest, usually dichotomous. For instance, important examples in clinical trials are the survival time and the disease free time (or recurrence time). In this context, the principal feature of the data is the possibility of censoring. Another important aspect of this type of data is whether or not the precise time of the end-point is known. Frequently, only the interval of occurrence of the event is known. For instance, patients are often examined periodically at fixed times but the event of interest may occur in between exams with no possibility to determine the exact occurrence time. Data in this form - known as interval-censored data, grouped or discrete lifetimes - require appropriate methods.

There is a vast literature on this topic and the articles by Lindsey & Ryan (1998) and Sun (1998) provide a good overview and several references. A variety of ways has been proposed for dealing with interval-censored data, as discussed in Cox (1972), Holford (1976), Tibshirani & Ciampi (1983), Finkelstein (1986), Farrington (1996), Huang, (1996), Huang & Rossini

(1997), and Goodall et al. (2004), among others. Obviously, all proposed methods of analysis have advantages and drawbacks, but a comparative work assessing the merits of each method is not available.

In the simplest situation in which only time is analyzed, the life table method can be used. However, frequently some covariates need to be incorporated into the analysis. For the regression case, the proportional hazard model is the most popular model, and when the proportional hazard assumption is not satisfied, the proportional odds model might be appropriate (see Huang & Rossini, 1997).

The treatment of censoring by the life-table method differs from the one using regression models. The discrete-time model is more general since it can incorporate several types of covariates. The life-table method is conceptually simple and available in several software packages, while the analysis of the discrete-time model is more complex and requires for instance knowledge of the generalized linear model. Moreover, the conclusions given by the two approaches may not be the same.

Thus, the comparison among those distinct types of analysis is an important issue in practice. Some questions arise: in which conditions would the life-table method be equivalent to the discrete-time models? Between the two regression models, which one should be chosen? Those issues have motivated the simulation study presented in Section 2.

In this paper we consider three types of analyses of interval-censored data: the life table method and two discrete-time regression models.

1.1 Life table method

The analysis of time until the event can be done by the traditional life table method (LTM) and the Mantel-Haenszel method can be applied for comparing the "survival" curves. The details on these methods can be found for example in Lawless (2003) and they are implemented in several commercial software packages or can be easily programmed.

1.2 Discrete-time models

In this section we consider two common discrete-time models: the proportional hazard and proportional odds models, referred to as discrete Cox model (DCM) and discrete logistic model (DLM), as detailed for instance in Lawless (2003, Chapter 7) and Collett (2003, Chapter 9). The score test related to these two models is given by Colosimo et al. (2000).

Let us consider the time T partitioned into k intervals ($I_i = [t_{i-1}, t_i)$, $i = 1, \ldots, k$), and let us assume that all censoring takes place at the end of the intervals. Let R_i be the risk set at time t_{i-1}, δ_{ij} an indicator variable (one if the event occurred for the jth individual within I_i and zero otherwise), x_j the vector of covariates, and $p_i(x_j) = \Pr(T_i \leq t_i | T_i \geq t_{i-1}, x_j)$, i.e. the

probability of failure of the jth individual in the interval I_i given that the failure did not occur before t_{i-1}. The likelihood function is given by

$$L = \prod_{i=1}^{k} \prod_{j \in R_i} [p_i(x_j)]^{\delta_{ij}} [1 - p_i(x_j)]^{1-\delta_{ij}}. \tag{1}$$

The form of $p_i(x_j)$ for DCM and DLM depends on the covariate effect (β) and the interval effect (γ) as follows. DCM is expressed as $p_i(x_j)$ $= 1 - [S_0(t_i)/S_0(t_{i-1})]^{\exp\{\beta' x_j\}} = 1 - \gamma_i^{\exp\{\beta' x_j\}}$, where $S_0(.)$ is the baseline survival function. After a simple algebraic manipulation, and calling $\gamma_i^* = \log(-\log(\gamma_i))$, the model DCM becomes

$$\log(-\log(1 - p_i(x_j))) = \gamma_i^* + \beta' x_j. \tag{2}$$

DLM is given by $p_i(x_j) = 1 - [1 + \gamma_i \exp\{\beta' x_j\}]^{-1}$. Taking the logit transformation and calling $\gamma_i^* = \log(\gamma_i)$, the model DLM can be written as

$$\log(p_i(x_j)/(1 - p_i(x_j))) = \gamma_i^* + \beta' x_j. \tag{3}$$

Note that those two models belong to the family of generalized linear models, and thus they can be fitted using the standard software packages, such as GLIM, SPlus and SAS, after an appropriate adjustment of entries of data for interval-censored data. Both have binomial error, and the link functions are complementary log-log and logit, respectively.

2   A Monte Carlo simulation study

We performed a simulation study for a comparison among the three types of analysis (LTM, DCM, DLM). In order to compare LTM with the two models (DCM and DLM), except group, we did not allow additional covariates, i.e. there was just one dichotomous covariate for models (2) and (3). We considered six time intervals, two groups, four sample sizes ($n = 60, 100, 200, 500$ for balanced designs, i.e. $n/2$ in each group), and three censoring proportions ($30\%, 40\%, 50\%$).

We evaluated the agreement with respect to the comparison between the two groups. In addition, we assessed the empirical power and the length of the confidence interval for the probability of failure in each time interval and the group parameter (β) of models (2) and (3). We also investigated the impact of misspecification of the regression model (i.e. we generated the data according to one model and proceeded to the analysis for the other one). The calculations were done in SPlus with 1000 simulations.

We generated the number of failures according to a binomial distribution with parameters n_i, the number of individuals at risk at the beginning of the time interval $I_i = [t_{i-1}, t_i)$, and $p_i(x_j) = \Pr(T_i \leq t_i | T_i \geq t_{i-1}, x_j)$. These probabilities were generated according to models (2) and (3) with

the following parameters: $\beta = 1$ and $\gamma_i^* = -3.5, -3, -2.5, -2, -1.5, -1$, i.e. an increasing effect of time on the risk of failure. The censoring was generated using the distribution $U(0,1)$. Finally, we applied the life table and the Mantel Haenszel methods, and for models (2) and (3) we tested $H_0 : \beta = 0$ and constructed the 95% confidence interval for β.
The main results are:

1. The agreement among the methods is always greater than 90%, regardless of the sample size and the censoring proportions, and it increases as the sample size increases.

2. As expected, as the sample size increases, the empirical power increases, but there is a reduction of power as the censoring proportion increases. The power is at least 66%, 55% and 40%, respectively for censoring proportions of 30%, 40% and 50%. For a fixed sample size and proportion of censoring, the power for the three types of analysis does not vary significantly.

3. As the sample size decreases and the censoring proportion increases, all the statistics for the length of the confidence interval for β increase. Smaller lengths are observed when the DCM is fitted, but the difference between the two models (DCM, DLM) nearly vanishes for large samples.

4. The impact of misspecifying the model is more noticeable when the DCM is the true model, confirming the higher accuracy of this model with respect to inference for the group parameter (β).

3   Concluding remarks

We have compared three types of analysis for interval-censored data (LTM, DCM, DLM) through a simulation study. An intriguing question is whether or not the complexity of DCM and DLM guarantees their superiority compared to LTM. Between the two regression models, the remaining question is which one should be used.
In the literature, LTM is recommended only for the case of large sample sizes. However, our results showed that the method works quite well even for small sample sizes (e.g. $n = 60$).
By comparing the three types of analysis we observed the effect of sample size and censoring proportion. Moreover, when comparing groups we concluded that the empirical powers were very similar; there was an excellent agreement in terms of deciding whether or not to reject the hypothesis of equal groups. In the comparison between the discrete-time models, there was an evidence of superiority of the DCM for the estimation of parameters and also for the wrong choice of the link function.
Finally, additional work is needed to cover other interesting situations, such as the inclusion of other types of covariates and unbalanced samples.

References

Collett, D. (2003). 2nd. ed. *Modelling Survival Data in Medical Research.* London: Chapman & Hall.

Colosimo, E.A., Chalita, L.V.A.S., Demetrio, C.G.B. (2000). Tests of proportional hazards and proportional odds models for grouped survival data. *Biometrics*, **56**, 1233-1240.

Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal and Statistics Society, B*, **34**, 187-220.

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

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

Goodall, R.L., Dunn, D.T., Babiker, A.G. (2004). Interval-censored survival time data: confidence intervals for non-parametric survivor function. *Statistics in Medicine*, **23**, 1131-1145.

Holford, T. (1976). Life tables with concomitant information. *Biometrics*, **32**, 587-597.

Huang, J. (1996). Efficient estimation for the proportional hazard model with interval censoring. *Annals of Statistics*, **24**, 540-568.

Huang, J., Rossini, A.J. (1997). Sieve estimation for the proportional odds failure-time regression model with interval censoring. *Journal of the American Statistical Association*, **92**, 960-967.

Lawless, J.F. (2003). *Statistical Models and Method for Lifetime Data.* New York: John Wiley.(1983). (1998).

Lindsey, J.C., Ryan, L.M. (1998). Tutorial in Biostatistics: methods for interval-censored data. *Statistics in Medicine*, **17**, 219-238.

Sun, J. (1998). Interval censoring. In: *Encyclopedia of Biostatistics*, 2090-2095, New York: John Wiley.

Tibshirani, R.J., Ciampi, A. (1983). A family of proportional- and additive-hazard models for survival data. *Biometrics*, **39**, 141-147.

Two-Stage Models to Control for Overdispersion in Longitudinal Count Data

Ali Reza Fotouhi[1]

[1] Dept. of Mathematics and Statistics University College of the Fraser Valley, 33844 King Road, Abbotsford, BC V2S 7M8, Canada, email:ali.fotouhi@ucfv.ca

Abstract: A family of two-stage models for longitudinal counts with different types of error terms is presented. These models account for overdispersion, serial correlation, and heteroscedasticity. The effects of omitted variables, link functions, and outliers are also investigated. The estimation approach is Markov Chain Monte Carlo within Gibs sampling. The proposed methods are applied to epileptic seizure counts data and illustrated in a simulation study.

1 Introduction

In applying standard Generalized Linear Models (GLMs) it is often found that the data exhibit greater variability than is predicted by the implicit mean-variance relation-ship. This phenomenon of overdispersion has been widely considered in the literature, particularly in relation to the Poisson distribution. In order to analyze overdisperd data we can broadly categorize the approaches into two groups. (i) Assume some more general form for the variance function with additional parameters and use quasi-likelihood approach. (ii) Assume a two-stage model for the response with the model parameter itself having some distribution. Thall and Vail (1990) have considered Generalized Estimating Equations (GEE) to model overdispersion in count data. Crouchley and Davies (1999) have shown that the GEE approach has limitations which restrict its usefulness. They have illustrated their theory by reanalyzing data on polyp counts.

In simple cases such as Poisson-Gamma models MLE approach is possible, although approximation methods often used when mixing distribution is not conjugate to the response distribution such as Poisson-Normal models. Aitkin (1999) has introduced an algorithm for Nonparametric Maximum Likelihood Estimation (NMLE) in GLMs with variance component structure. Another approach is a fully Bayes approach with the additional structure of a prior distribution on all the model parameters. Fotouhi (2003) has shown that this approach performs very well in fitting multi-level models

especially for two-level models for analyzing longitudinal data. This approach will be used in this paper.

The principal objective of this paper is to explain the sources of overdispersion in longitudinal count data. We will specially show that the way of introducing the random effects into the linear predictor is essential to overcome the problem of overdispersion.

2   Data

We report analysis of two well known data sets. The first one is data on epileptic seizure count arising in a study of progabide as an adjuvant antiepileptic chemotherapy. The data are from a clinical trial of 59 epileptics reported and analyzed by Thall and Vail (1990). The second data set is from a 4-year randomized double-blind trial of treatments (58 patients) to reduce rectal polyps in sufferers of familial polyposis. The data are reported and analyzed by Crouchley and Davies (1999). The seizure counts exhibit a high degree of extra-poison variation for total data, placebo and progabide groups, baseline, and each visit. Moreover the seizure counts exhibit heteroscedastic overdispersion across visit and across treatment group. Almost the same patterns could be found in polyp data.

3   Theory

Assume that, conditional on error term ε_{it}, Y_{it} is distributed as Poisson with mean $\lambda_{it} = \mu_{it}\psi_{it}$ where $\psi_{it} = \exp(\varepsilon_{it})$ and $\mu_{it} = \exp(\eta_{it})$. The second term in marginal variance of Y_{it}, $Var(Y_{it}) = \mu_{it}E(\psi_{it}) + \mu_{it}^2 Var(\psi_{it})$, shows overdispersion. The dependency of this term on time indicates the heteroscedasticity of overdispersion. To overcome the problem of overdispersion we decompose ε_{it} into three components, random effects γ_i, serial correlation ξ_t and measurement error, δ_{it} (see Diggle et al. (1994). For epileptic data the linear predictor, including all three error terms γ_i, ξ_t, and δ_{it}, may be of the form

$$
\begin{aligned}
\lambda_{it} \quad = \quad & \exp[\beta_0 + \beta_1(\log Age - mean(\log Age)) + \beta_2(\log(Base/4) - mean(\log(Base/4))) \\
& + \beta_3(Trt. - mean(Trt.)) + \beta_4(Visit - mean(Visit)) + \beta_5(Trt. \times \log(Base/4) \\
& - mean(Trt. \times \log(Base/4))) + \gamma_i + \xi_t + \delta_{it}]
\end{aligned}
$$

where $Visit$ is binary indicator for the fourth clinic visit and $Trt.$ is 0 for placebo and 1 for progabide.

To use Bayesian inference Using Gibbs Sampling (BUGS) we assume specific parametric priors for γ_i, ξ_t, and δ_{it}. Let $\gamma_i \sim NID(0, \sigma_\gamma)$, $\xi_t \sim MND(0, V)$, $\delta_{it} \sim NID(0, \sigma_\delta)$. We assume non-informative priors with extremely small precision for the structural parameters β_0, β_1, β_2, β_3, β_4, β_5, ($\beta_j \sim N(0, 10000)$). We also assume non-informative prior with mean 1 and variance 1000 for the precisions of the error terms, i.e. $\frac{1}{\sigma_\gamma} \sim$ Gamma $(0.001, 0.001)$ and $\frac{1}{\sigma_\delta} \sim$ Gamma $(0.001, 0.001)$. The prior distribution for

covariance matrix V is assumed to be Wishart with appropriate parameters.

We use three model checking criteria in both application and simulation study. they are Deviance, Variance Inflation Factor (VIF), and global goodness-of-fit tests based on Bayesian probability ($p - value$). We also check if the estimated model is consistent to the data.

4   Application and Simulation

We have fitted the proposed models to epileptic and polyp data and have done some simulations. We report only some of our findings in table 1. We have used BUGS program and for all models, a burn-in of 3000 iterations was followed by a further 6000 iterations.

Table 1 shows that for the model with no error term, VIF is 4.454, which shows the existence of overdispersion. Changing the link function and deleting two outliers does not change the VIF significantly but change the deviance. According to the global goodness-of-fit tests based on Bayesian probability ($p - value$) none of the models are fitted significantly. The threshold for no error term model is not consistent to the data.

Considering serial correlations among repeated counts within patients by introducing a multivariate Normal random vector ξ in the linear predictor does not reduce the VIF and the deviance.

The random effects model γ_i performs better than the no error term model. The VIF and deviance reduce substantially to 1.841 and 1221 respectively but VIF is still significantly larger than 1. The standard error of the individual specific error, σ_γ, is estimated 0.538(0.064) which is significantly different from $zero$. This shows that the heterogeneity across individuals is captured.

Table 1 shows that all models having measurement error, δ_{it}, are fitted perfectly well. The standard error of the measurement error is estimated significantly different from $zero$ in all these models. The VIF and deviance for these models are minimum comparing to the models having the same specifications but not including δ_{it}. The VIF for these models is close to 1, showing that overdispersion is completely captured. The Bayesian $p - values$ for the models containing δ_{it} suggest that the observed Pearson χ^2 statistic is consistent with the value expected from a random sample of 59×4 from a Poisson distribution. That is, there is no evidence against the assumption about the structure of the underlying linear predictor. Our best fitted models are the measurement error model ,δ_{it}, and the model containing both random effects, γ_i, and measurement error, δ_{it} . The mentioned three criteria do not distinguish these two models. But the threshold is 54.1 for the first model and 72.6 for the second model. The model containing both the random effects and measurement error is then more consistent to the data, since the patient with baseline 67 has been substantially recovered after receiving treatment.

Thall and Vail (1990) have fitted several models to the epileptic seizure data. The model with both individual random effects and independent time random effects has been introduced as the best model. We have calculated the threshold for this model and is 60.8. This threshold is not very consistent to the data since two patients with baselines 67 and 76 have been greatly recovered after receiving treatment. Our model with $\gamma_i + \xi$ is equivalent to their best model for which $VIF = 1.975$ showing that overdispersion is not completely captured.

Comparing the fitted models, we observe a systematic reduction of the standard deviation of the parameter estimate with increasing VIF. This shows that lack of controlling the overdispersion arising from the omitted variables may overstate the significance of explanatory variables. We have also fitted several models and have investigated the effect of the initial conditions problem (Fotouhi (1997)) on overdispersion. Even if initial conditions are treated correctly we still need a proper consideration of the error terms.

The second application is applying the proposed models to analyze polyp data reanalyzed by Crouchley and Davies (1999). They have shown that random effects model is more appropriate than GEE approach for assessing the treatment effects for these data. We have calculated the VIF for their model and that is 8.35. We have shown that the model including measurement error, δ_{it} performs better in capturing overdispersion with $VIF = 5.29$ and produces more consistent thresholds for assessing the treatment effects. Non of these models could capture the overdispersion completely. Perhaps the overdispersion is not due to omitted variables. The overdispersion in epileptic data was due to omitted variables and controlled by proper consideration of the error term. Our simulation study based on epileptic data shows the same patterns obtained from application of the same models to epileptic data. VIF, $deviance$, and $p - value$ for measurement error model are 1.034, 1398, and 0.490 respectively. While for random effects model are 9.931, 3367, and 0 respectively. We conclude that if the data are produced by a process affected by measurement error then the random effect model is not able to capture the overdispersion.

5   Concluding remarks

We introduced some models with different types of error terms in their linear predictor to control for omitted variables and consequently to control for overdispersion in longitudinal count data analysis. We have shown, through application to epileptic seizure and polyp data and simulation, that the type of the error term is important to overcome the problem of overdispersion. We have also shown that the link function and the outliers are also important factors. As expected, the standard error of estimate increases as VIF decreases.

TABLE 1. Parameter estimates and goodness of fit criteria from fitting model 2 with different types of error term. Bold figures shown are standard deviations of estimates.

$Model$	No $Error$	ξ	γ_i	δ_{it}	$\xi + \delta_{it}$	$\gamma_i + \delta_{it}$	$\gamma_i + \xi$	$\gamma_i + \xi$ $+\delta_{it}$
$Int.$	1.686	2.224	1.620	1.561	1.940	1.567	-3.231	4.948
	0.032	**0.412**	**0.082**	**0.052**	**0.533**	**0.074**	**0.457**	**0.394**
Age	0.889	0.887	0.483	0.578	0.564	0.493	0.451	0.486
	0.117	**0.116**	**0.375**	**0.241**	**0.240**	**0.363**	**0.348**	**0.372**
$Base$	0.947	0.951	0.882	0.899	0.917	0.914	0.906	0.906
	0.044	**0.042**	**0.124**	**0.081**	**0.085**	**0.133**	**0.112**	**0.130**
$Trt.$	-1.343	-1.330	-0.896	-0.982	-0.947	-0.864	-0.816	-0.879
	0.158	**0.145**	**0.351**	**0.254**	**0.270**	**0.430**	**0.308**	**0.374**
$Visit$	-0.160	1.364	-0.160	-0.093	3.663	-0.103	0.957	-2.203
	0.054	**1.156**	**0.055**	**0.114**	**1.720**	**0.087**	**1.280**	**0.650**
BT	0.563	0.558	0.320	0.377	0.360	0.298	0.280	0.313
	0.064	**0.058**	**0.174**	**0.118**	**0.132**	**0.222**	**0.149**	**0.189**
σ_γ	–	–	0.539	–	–	0.498	0.520	1.418
	–	–	**0.064**	–	–	**0.070**	**0.891**	**0.069**
σ_δ	–	–	–	0.594	0.596	0.360	–	0.364
	–	–	–	**0.046**	**0.435**	**0.043**	–	**1.789**
VIF	4.454	4.804	1.841	1.077	1.143	1.063	1.975	1.120
$Dev.$	1641	1641	1221	1037	1035	1038	1222	1035
PV	0	0	0	0.380	0.375	0.416	0	0.424

References

Aitkin, M. (1999). A General Maximum Likelihood Analysis of Variance Components in Generalized Linear Models. *Biometrics* **55**, 117-128.

Crouchley, R. and Davies, R. B. (1999) A comparison of population average and random effect models for the analysis of longitudinal count data with base-line information. *JRSS A*, **162**, 3, 331-347.

Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1994) Analysis of longitudinal Data. Oxford: Clarendon.

Fotouhi, A. R. (1997). Longitudinal Data Analysis: The Initial Conditions Problem in Random Effects Modelling. Ph.D. Thesis in Centre for Applied Statistics Lancaster University.

Fotouhi, A. (2003) Comparison of estimation procedures for multilevel models. *Journal of the statistical software*, **8**, 9.

Thall, P. F. and Vail, S. C. (1990) Some covariance models for longitudinal count data with overdispersion. *Biometrics*, **46**, 657-671.

Multilevel Logit Models: A Comparison of Estimation Procedures

Ali Reza Fotouhi[1]

[1] Dept. of Mathematics and Statistics University College of the Fraser Valley, 33844 King Road, Abbotsford, BC V2S 7M8, Canada, email:ali.fotouhi@ucfv.ca

Abstract: We introduce Multilevel Logit Models and discuss the estimation procedures that may be used to fit these models. We apply the proposed procedures to three-level binary data generated in a simulation study. We compare the procedures by two criteria, Bias and efficiency. We find that the estimates of the fixed effects and variance components are substantially and significantly biased using Longford's Approximation and Goldstein's Generalized Least Squares approaches by two software packages VARCL and ML3. These biases could be removed by using Markov Chain Monte Carlo (MCMC) using Gibbs sampling or Nonparametric Maximum Likelihood (NPML) approach. The Gaussian Quadrature (GQ) approach, even with small number of mass points results in consistent estimates but computationally problematic.

1 Introduction

In multilevel data, the observations within the same group are more likely to be correlated than the observations from different groups. The correlations from all levels should be taken into account and ignoring any one of them may lead to inconsistent estimates and misleading inferences. A well known method of representing this common variation is by adding a common unobserved random effect to the linear predictor for each lower level unit in the same upper level unit. If the distribution of this random effects is conjugate to the distribution of the responses, then maximum likelihood is straightforward. Otherwise the likelihood function does not have a closed form and we need an approach to deal with the integration problem. Some approaches to solve the integrals are:(a) The likelihood can be integrated numerically using Gaussian Quadrature (GQ) points. (b) The log likelihood function can be approximated by a second order Taylor series expansion. (c) A fully Bayesian approach can be used with the additional structure of a prior distribution on all the model parameters. The Markov Chain Monte Carlo (MCMC) methods can be used to obtain marginal posterior distributions of the parameters.

In these three approaches we assume a specific parametric form of the mixing distribution of the unobserved random effects. Davies (1987) has

shown that the parameter estimation is sensible to the choice of the mixing distribution. This problem can be solved by Nonparametric Maximum Likelihood (NPML) estimation on mixing distribution on a finite number of mass points. This approach is used by Aitkin (1999) for fitting two-level data.

Very little work has been done on using and comparing the four mentioned approaches, GQ, Taylor series, MCMC, and NPML in analyzing multi level data. The purpose of this paper is to model a multilevel binary data in a general form and explain, apply and compare the above approaches through simulation study. Our analysis focuses on bias and efficiency of estimates produced by the mentioned approaches. However the results will compare some software in fitting multilevel models.

2   Model and Estimation Approaches

Following Goldstein (1991) a multilevel logit model is of the form,

$$logit(\mu) = \eta = \mathbf{X}\beta + \mathbf{Z}\mathbf{u}$$

where $\mu_i = P_r(Y_i = 1|\beta, \mathbf{\Omega}, \mathbf{X}, \mathbf{Z})$; for $i = 1, ..., N$ and η is a conditional linear predictor. We assume that the random effects from different units are mutually independent with mean $\mathbf{0}$ and $Var(\mathbf{u}_i) = \mathbf{\Omega}_i$. We then have $Var(\mathbf{u}) = \mathbf{\Omega}$ and $\mathbf{\Omega} = diag_L[\mathbf{I}_{g_i} \otimes \mathbf{\Omega}_i]$.

To compare approaches we consider a three level logit model with one random effect at each of the second and third levels. If we consider one explanatory variable at each level then the above model reduces to

$$L(\beta, \mathbf{\Omega}) = \prod_{i=1}^{L} \int_{-\infty}^{+\infty} \left(\prod_{j=1}^{n_i} \int_{-\infty}^{+\infty} \left(\prod_{k=1}^{n_{ij}} \mu_{ijk} \right) g_1(u_{ij}) du_{ij} \right) \times g_2(u_i) du_i$$

$$\mu_{ijk} = \frac{\exp\left[(\beta_1 x_{ijk} + \beta_2 x_{ij} + \beta_3 x_i + u_i + u_{ij}) y_{ijk} \right]}{1 + \exp\left[\beta_1 x_{ijk} + \beta_2 x_{ij} + \beta_3 x_i + u_i + u_{ij} \right]}$$

where x_{ijk}, x_{ij}, and x_i are the explanatory variables in levels one, two, and three respectively. u_{ij} and u_i are the random effects with means zero and standard errors σ_1, σ_2 and density g_1, g_2 related to second and third levels respectively. y_{ijk} is the response for the kth individual in the jth unit of level two and i^{th} unit of level one. $\beta_1, \beta_2, \beta_3$ are the fixed effects of x_{ijk}, x_{ij}, and x_i. Here we need to calculate one dimensional integral.

Longford (1988) has proposed an approximation to this likelihood function. The approximation relies on a second order Taylor expansion of the logarithm of the conditional likelihood about $\mathbf{u} = \mathbf{0}$. Longford (1988) has implemented this estimation strategy in the software package VARCL. This method provides the basis for a Fisher scoring procedure which can be applied alternately to β and $\mathbf{\Omega}$. Although, Longford's approximation has

solved the problem of high dimensionality of the integrals for some models but care should be taken in applying this approximation. Since the true likelihood function is not maximized and the remainder of the Taylor expansion is not controlled the parameter estimate may by biased. Even if all the necessary conditions needed to write the Taylor series of the likelihood function are attained we need to control the remainder of the estimation of the likelihood function by its finite Taylor series. The same problem appears when we use the method proposed by Goldstein (1991) used in ML3. We will compare these two approaches with the three well known approaches MCMC, NPML, and GQ explained in introduction.

3   Simulation Study

In empirical study, unlike simulation study, since the true value of the parameters are not known we can never be certain if the results of empirical work are accurate and so we may have misleading comparisons of underlying approaches. For comparisons of estimation procedures we followed the simulation's structure proposed by Rodriguez and Goldman (1995). They have simulated data sets using the same hierarchial structure as one of the Guatemalan data sets analyzed by Pebley and Goldman (1992).
Consider 20 units in each level of the three-level model introduced in section 2. Suppose that x_{ijk}, x_{ij}, and x_i are dummy variables in fully balanced design, so the covariates are independent and each of the eight combinations of values occur equally often. the fixed effects $\beta_1, \beta_2, \beta_3$ are set to be one. The random effects u_{ij} and u_i are generated from independent normal distributions with means *zero* and variances 1.0 and 0.16. Tables 1 reports values of the estimated fixed effects and the estimated standard errors of the random effects averaged over the 100 simulations when the variance of random effects is 1.
The results from Rodriguez and Goldman (1995), reported in table 1, show large significant biases for all parameters. When they used VARCL software, except the fixed effect at third level, all the other estimates are significantly biased. Their performance in ML3 results in substantial significant biases especially for the standard error of the random effect at second level which are 89.7 and 72.2 percent using linear and quadratic approximations respectively. They have not reported the standard deviations of the estimates to check if the biases are statistically significant.
To implement the GQ approach we have used the subroutine BCONF from Fortran Power Station 4.0 software to maximize the likelihood function. Table 1 shows that none of the biases are statistically significant. We found that this approach behave poorly in estimating the standard error of the third level and is computationally problematic.
To apply the NPML approach we have used the subroutine LCONF from Fortran Power Station 4.0 software to maximize the likelihood function.

Table 1 shows that the results from the performance of the NPML approach with 3 mass points are better than the results from the GQ approach in estimating the standard errors of the random effects. Non of the biases from this approach are significant.

To apply the MCMC approach using Gibbs sampling we have used BUGS software. It is assumed that the prior distributions of u_i and u_{ij} to be normal with means 0 and standard errors σ_1 and σ_2 respectively. $\beta_1, \beta_2, \beta_3$ have non-informative normal prior with mean 0 and standard error 1000, σ_1 and σ_2 have non-informative gamma prior with mean 1 and variance 1000. In order to get over the influence of the initial values we have performed 500 iterations of the Gibbs sampler and then have updated another 1000 iterations to estimate the parameters. Table 1 shows that this approach performs excellent with at most 2.8% bias for the fixed effect at the second level. The standard deviation of estimates are small and none of the biases are statistically significant. Table 1 shows that the MCMC approach results in very small MSE.

Further investigations showed that when the variances of the random effects are small, i.e. $\sigma_1^2 = \sigma_2^2 = 0.16$, non of the estimates are significantly biased. Using GQ or NPML results in large absolute biases for the standard errors of the random effects but are not statistically significant. VARCL and BUGS perform almost the same but with less biases using BUGS.

4   Conclusions

In this paper we reviewed the procedures that may be applied to fit multi-level logit models and compared these approaches through simulation study. We showed that the substantial significant biases coming from VARCL and ML3 can be vanished by applying the MCMC method using Gibbs sampling. The efficiency of the MCMC approach is considerably high and recommended if we assume a parametric distributions for the random effects. If there is not such prior information, the NPML approach is recommended. Our simulation study shows that this approach performs better than VARCL and ML3.

References

Aitkin, M. (1999). A General Maximum Likelihood Analysis of Variance Components in Generalized Linear Models. *Biometrics* **55**, 117-128.

Davies, R. B. (1987). Mass Point Methods for Dealing with Nuisance Parameters in Longitudinal Studies. In: R. Crouchley (ED.) Longitudinal Data analysis (Avebury, 1987) 88-109.

TABLE 1. Simulation results for large error variance. The figures in the parentheses are the standard deviations of estimates. Bold figures are MSE of the estimates. * Significantly biased estimates

Approach	$\beta_1 = 1$	$\beta_2 = 1$	$\beta_3 = 1$	$\sigma_1 = 1$	$\sigma_2 = 1$
	0.756*	0.775*	0.906	0.801*	0.749*
VARCL	(0.062)	(0.089)	(0.378)	(0.044)	(0.115)
	0.063	**0.059**	**0.152**	**0.042**	**0.076**
	1.149	1.017	1.035	0.957	1.994
GQ	(0.408)	(0.378)	(0.674)	(0.425)	(1.073)
	0.189	**0.143**	**0.456**	**0.182**	**2.139**
	1.003	0.972	0.756	1.350	1.243
NPML	(0.063)	(0.155)	(0.467)	(0.244)	(0.315)
	0.004	**0.025**	**0.278**	**0.182**	**0.158**
	0.992	0.972	1.010	1.000	0.997
MCMC	(0.115)	(0.118)	(0.350)	(0.062)	(0.199)
	0.013	**0.015**	**0.123**	**0.009**	**0.040**
ML3-Linear	0.738	0.74	0.771	0.103	0.732
ML3-Quadratic	0.854	0.860	0.910	0.278	0.764

Goldstein, H. 1991. Nonlinear Multilevel Models with an Application to Discrete Response Data. *Biometrika*, **78**, 1, 45-51.

Longford, N. T. (1988). VARCL: Software for Variance Component Analysis of Data With Hierarchically Nested Random Effects (Maximum Likelihood). Princeton, N. J., Educational Testing Service.

Pebley, A. R. and Goldman, N. (1992). Family, Community, Ethnic Identity and the use of Formal Health Care Services in Guatemala. OPR Working Paper 92-102. Office of Population Research, Princeton.

Rodriguez, G. and Goldman, N. (1995). An Assessment of Estimation Procedures for Multilevel Models with Binary Responses. *Journal of the Royal Statistical Society, Series A* **158**, part 1, 73-89.

Seasonal Variation in Death Counts: P-Spline Smoothing in the Presence of Overdispersion

Jutta Gampe[1], Roland Rau[1]

[1] Max Planck Institute for Demographic Research
Konrad-Zuse-Str. 1, 18057 Rostock, Germany
email: gampe@demogr.mpg.de, rau@demogr.mpg.de

Abstract: Overdispersion may have a considerable influence on smoothing results if the extra variability is not accounted for in the model. We propose a two-stage strategy for the estimation of the overdispersion and smoothing parameters in a negative binomial varying-coefficient model.

1 Introduction

Death does not strike uniformly over the year. On the Northern hemisphere deaths typically peak in winter whereas mortality is lowest late in summer (July-September). These seasonal fluctuations are a persistent phenomenon in most populations and they follow a sinusoidal shape rather closely. Climatic conditions — mainly temperature — shape the seasonal variation in risks of death, however, social factors modulate seasonal mortality patterns as well. Mortality patterns have changed considerably over the last decades, the most striking development being the dramatic and unprecedented progress against mortality at advanced ages. Whether seasonal fluctuations have undergone similar changes and whether some age-groups or causes of death benefitted more from general improvements in living conditions and medical progress than others is not yet fully known.

2 Data

The data set used in this study was derived from the *"Multiple Cause of Death" public use files* published by the US Center for Disease Control and Prevention (CDC) for the years 1959–1998. The data consist of more than 77 Mio. individual deaths records. Each record contains information on the sex of the individual, month and year of death, age at death, and cause of death. The emphasis here is on adult and especially old-age mortality and therefore only deaths that occurred at ages 50 and higher were included.
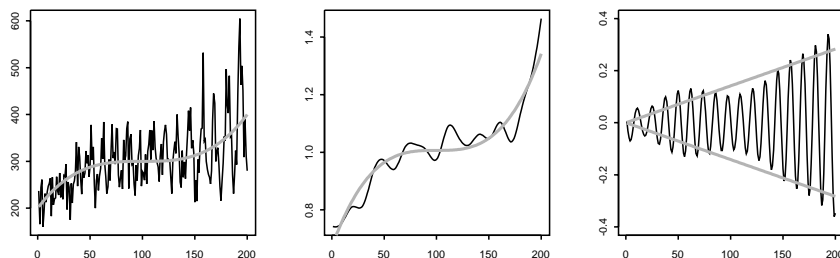
FIGURE 1. Smoothing results when overdispersion is ignored. Data (left) were simulated from a Negative Binomial distribution but the model was fitted under a Poisson assumption. Additive trend (middle) and amplitude function of the seasonal component (right) are both undersmoothed. (True values in gray, fitted values in black.)

The purpose of the study was to find out whether and how seasonal variation in mortality had changed over the observation period for different age-groups and different causes of death.

3   Modelling Changing Seasonal Variation

The overall trend in the number of deaths is determined by changes in age-group sizes and changing mortality risks over time and should be modelled flexibly. Additionally we want to obtain a flexible and data-driven estimate for potential changes in seasonal mortality fluctuations.

3.1   Model and *P*-Spline Smoothing

We denote the monthly numbers of deaths (for a specific cause of death and age-category) by Y_t, $t = 1, \ldots, T = 480$ ($\hat{=}$Jan '59, ..., Dec '98). We start by assuming that the Y_t are independently Poisson distributed with a log-link and the mean μ_t specified as

$$\ln \mu_t = \alpha_0 + f_0(t) + \sum_{l=1}^{L} \left\{ f_l^{\sin}(t) \, \sin(\frac{2\pi\, l}{12}\, t) + f_l^{\cos}(t) \, \cos(\frac{2\pi\, l}{12}\, t) \right\}. \quad (1)$$

Both the additive trend term $f_0(t)$ and the amplitude modulating functions $f_l^{\sin}(t)$ and $f_l^{\cos}(t)$ are assumed to be smoothly varying functions over time t. The most simple seasonal model would only fit one sine-cosine term ($L = 1$), by adding more components more complex cyclic patterns could be captured. Model (1) is a varying-coefficient model (Hastie and Tibshirani,

1993) which, as demonstrated by Eilers and Marx (2002), can be conveniently fit using P-splines. Each smooth model component is expanded using a moderately large B-Spline basis and smoothness is controlled by penalizing the spline-coefficients by a difference penalty (Eilers and Marx, 1996). The optimal amount of smoothing can be determined by minimizing an information criterion, like AIC, over a grid of values for the smoothing parameter λ. For large models with several functions to be smoothed Eilers and Marx (2002) suggest a multi-dimensional grid-search to determine the optimal combination of smoothing parameters.

3.2   The Impact of Overdispersion

Clearly there is unobserved heterogeneity in these data. The month index is only a proxy for the actually prevailing weather conditions, and individuals, even for narrow age categories, have different susceptibility to death. Both features are well known sources of overdispersion (Cameron and Trivedi, 1998; Barron, 1992). The effect of overdispersion on smoothing methods can be considerable and is depicted in Figure 1. Extra variation that is not allowed for by the Poisson model is distributed over the smooth model components leading to serious undersmoothing of the target functions. This phenomenon corresponds to the similar effect that arises when correlated data are smoothed under independence assumptions.

3.3   Smoothing Parameter Selection

A simple and common extension for overdispersed count data is the Negative Binomial (NB) distribution (Lawless, 1987), arising from a Gamma-Poisson mixture. For a fixed value of the variance τ^2 of the mixing Γ-distribution (with mean 1), the NB is an exponential family and we thus still operate in the GLM framework. Therefore, for a given amount of overdispersion τ^2, we may determine the values of the smoothing parameters as in the Poisson case. An optimal procedure though has to determine which portion of the variation in the data can be attributed to overdispersion and which is due to the structural components in the model. To resolve this question we propose the following two-stage strategy.

- Fix a grid of values for the overdispersion parameter, i.e. the variance of the Γ-distribution:
$$\tau_1^2, \ldots, \tau_M^2.$$

- For each of these (fixed) values τ_m^2 $(m = 1, \ldots, M)$ minimize the AIC to obtain the optimal smoothing parameters $(\lambda_1^m, \ldots, \lambda_C^m)$, where C is the number of components to be smoothed in (1).
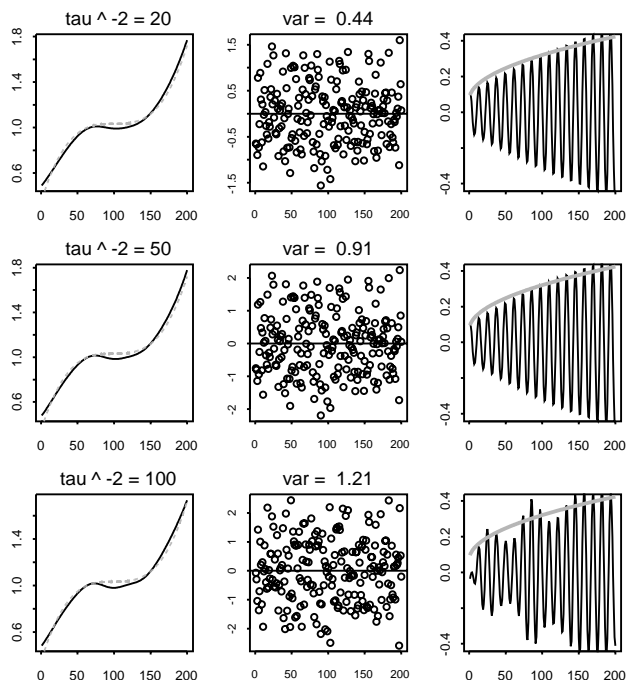
FIGURE 2. Results from a simulation study applying the two-stage smoothing strategy. Each row shows the true (dashed) and estimated (solid) trend function (left), the Pearson residuals and their variance (middle), the seasonal component (estimate and true amplitude; right) for a fixed value of overdispersion τ^2. The true value in this case was $\tau^2 = 1/50$.

- For these smoothing parameters calculate the Pearson residuals according to the NB model currently under consideration (i.e. the fixed value τ_m^2)

$$p_t = \frac{y_t - \hat{\mu}_t}{\sqrt{\hat{\omega}_t}} \qquad \hat{\omega}_t = \hat{\mu}_t + \tau_m^2 \hat{\mu}_t^2$$

- Choose as the final model the combination $(\tau_{m^*}^2; \lambda_1^{m^*}, \ldots, \lambda_C^{m^*})$ for which the variance of the Pearson residuals is ≈ 1.

4   Results

Figure 2 shows results obtained by this procedure from a larger simulation study, demonstrating the interplay between overdispersion and opti-
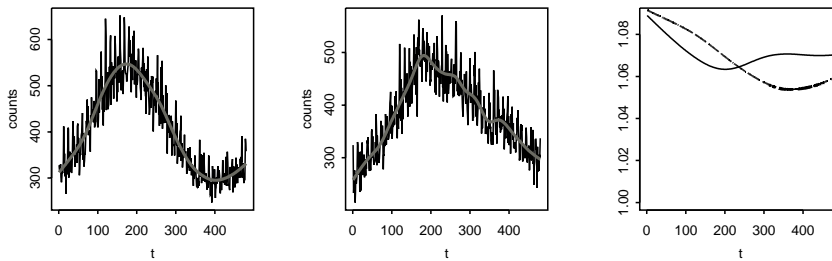
FIGURE 3. Deaths due to cirrhosis, men, ages 50-59 (left), ages 60-69 (middle). Right: Modifying functions of seasonal amplitudes. Dashed: ages 50-59, solid: ages 60-69.

mal smoothing. In Figure 3 the estimated functions for male deaths due to cirrhosis in two different age groups (50–59 and 60–69) are compared.

References

Barron D.N. (1992). The Analysis of Count Data: Overdispersion and Autocorrelation. *Sociological Methodology*, 179–220.

Cameron, A.C. and P.K. Trivedi](1998). *Regression Analysis of Count Data*. Econometric Society Monographs. Cambridge University Press.

Hastie, T. and R. Tibshirani (1993). Varying-coefficient Models. *Journal of the Royal Statistical Society B*, **55**, 757–796.

Eilers, P. H. and B. D. Marx (1996). Flexible smoothing using B-splines and penalized likelihood. *Statistical Science*, **11**, 89–121.

Eilers, P. H. and B. D. Marx (2002). Generalized linear additive smooth structures. *Journal of Computational and Graphical Statistics*, **11**, 758—783.

Lawless J.F. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics*, **15**, 209–225.

Power-Divergence Goodness-of-Fit Statistics: Small Sample Behavior in One Way Multinomials and Applications to Multinomial Processing Tree (MPT) Models

Vicente Núñez-Antón[1] and Miguel A. García-Pérez[2]

[1] Departamento de Econometría y Estadística, Universidad del País Vasco, Av. del Lehendakari Aguirre 83, 48015 Bilbao, Spain, e-mail: `etpnuanv@bs.ehu.es`

[2] Departamento de Metododología, Facultad de Psicología, Universidad Complutense, Campus de Somosaguas, 28223 Madrid, Spain

Abstract: The small-sample behavior of power-divergence goodness-of-fit statistics with composite hypotheses is evaluated in multinomial models of up to five cells and up to three parameters. These models were based on a class of cognitive models called *multinomial processing tree* (MPT) models, that are characterized for being simple and substantively motivated statistical models than can be applied to categorical data. They are used as data-analysis tools for measuring underlying or latent cognitive capacities and as simple models for representing and testing competing psychological theories. The performance of these tests was assessed by comparing asymptotic sizes with exact sizes obtained by enumeration. This paper addresses all combinations of power-divergence estimates of indices $\nu = \{-1/2, 0, 1/3, 1/2, 2/3, 1, 3/2\}$ and statistics of indices $\lambda = \{-1/2, 0, 1/3, 1/2, 2/3, 1, 3/2\}$. Exact conditions are given under which the asymptotic approximation is sufficiently accurate, by the criterion that the average exact size is no larger than $\pm 10\%$ of the asymptotic test size.

Keywords: One-way multinomial; Goodness-of-fit; Power divergence statistic; MPT models; Parameter estimation; Composite hypothesis; Exact test size.

1 Introduction and Method

Let $\boldsymbol{O} = (O_1, O_2, \ldots, O_k)$ with $k > 1$, $\sum_{i=1}^{k} O_i = n$ and $O_i \geq 0$ (for all $1 \leq i \leq k$) be the empirical distribution of n observations into k classes, and let $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_k) \in (0,1)^k$, with $\sum_{i=1}^{k} \pi_i = 1$ be a discrete distribution describing the probability of an observation's falling into each class. Then,

$$P(\boldsymbol{O}; \boldsymbol{\pi}) = n! \prod_{i=1}^{k} \frac{\pi_i^{O_i}}{O_i!} \tag{1}$$

is the probability of $\boldsymbol{O}$ under $\boldsymbol{\pi}$. Many goodness-of-fit problems involve parametric models in which $\boldsymbol{\pi}$ is merely assumed to belong in a set Π_0

of distributions whose elements are functionally dependent on some parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s) \in \mathbb{R}^s$ with $s \geq 1$. In other words, the model states that $\boldsymbol{\pi} \in \Pi_0$ with $\Pi_0 = \{\boldsymbol{\pi} \in (0,1)^k : \boldsymbol{\pi} = \boldsymbol{f}(\boldsymbol{\theta})\}$, where $\boldsymbol{f}(\boldsymbol{\theta}) = (f_1(\boldsymbol{\theta}), \ldots, f_k(\boldsymbol{\theta})) \in (0,1)^k$ and $\sum_{i=1}^{k} f_i(\boldsymbol{\theta}) = 1$. Testing the fit of the model Π_0 to the data $\boldsymbol{O}$, i.e., testing the null hypothesis $H_0 : \boldsymbol{\pi} \in \Pi_0$, requires estimating $\boldsymbol{\theta}$. Provided and efficient method is used to determine $\hat{\boldsymbol{\pi}} = \boldsymbol{f}(\hat{\boldsymbol{\theta}}) \in \Pi_0$ that is most consistent with $\boldsymbol{O}$, the power divergence statistic

$$2\boldsymbol{I}^\lambda(\boldsymbol{O} : \hat{\boldsymbol{e}}) = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{k} O_i \left\{ \left(\frac{O_i}{\hat{e}_i}\right)^\lambda - 1 \right\} \tag{2}$$

with $\lambda \in \mathbb{R}$, $\hat{e}_i = n\hat{\pi}_i = nf_i(\hat{\boldsymbol{\theta}})$, and $s < k-1$ is asymptotically distributed as a χ^2 r.v. on $k - s - 1$ degrees of freedom (Cressie and Read, 1984).

This asymptotic result may not provide an accurate approximation in the typical small-sample case and, there are reasons to believe that it will fail to do so: in the case of simple null hypotheses (i.e., with a completely specified $\boldsymbol{\pi}$), an analogous asymptotic result often yields inaccurate test sizes when at least one expectation is small (García-Pérez and Núñez-Antón, 2001), and some expectations are likely to be small with composite hypotheses. The accuracy of the asymptotic approximation in one-way multinomials with composite hypotheses has never been studied extensively (Larntz, 1978; Riefer and Batchelder, 1991; and García-Pérez, 1994). This paper evaluates systematically the small-sample accuracy of the asymptotic approximation for a broad set of conditions involving a range of one-way multinomial models with up to three parameters (i.e., $1 \leq s \leq 3$) and up to five cells (i.e., $3 \leq k \leq 5$), as a function of the power-divergence index λ.

These models were based on a class of cognitive models called *multinomial processing tree* (MPT) models (Riefer and Batchelder, 1988; or Batchelder and Riefer, 1999), that are characterized for being simple and substantively motivated statistical models than can be applied to categorical data. These models are used as data-analysis tools for measuring underlying or latent cognitive capacities and as simple models for representing and testing competing psychological theories. Based on the motivation of the MPT models, we have included in the study one-parameter models with $k = 3, 4, 5$ cells, a two-parameter model with $k = 4$ cells, and a three parameter model with $k = 5$ cells. In all cases the parameter space is $\Omega = (0,1)^s$. The one-parameter models for each k arise from the expansion of $[\theta + (1-\theta)]^m$, $1 \leq m \leq k-1$ and, then, the various π_i are polynomials in θ ranging from first degree up to $(k-1)$-th degree (see Figure 1). We consider sample sizes $n = 5k, 10k, 20k, 40k$. The study covers power-divergence statistics of indices $\lambda = -1/2, 0, 1/3, 1/2, 2/3, 1$, and $3/2$; in each case, parameter estimates were obtained by minimizing power-divergence measures of indices $\nu = -1/2, 0, 1/3, 1/2, 2/3, 1$, and $3/2$ also. We included all cases of matched statistics and estimate indices ($\lambda = \nu$, as advocated by Read and Cressie, 1988) and all combinations of mismatched indices ($\lambda \neq \nu$, as

shown to behave better on occasions by García-Pérez, 1994).

The exact distribution function was obtained using the procedure described in García-Pérez and Núñez-Antón (2004). Multinomial probabilities $P(\boldsymbol{X};\hat{\boldsymbol{\pi}})$ were obtained with the algorithm in García-Pérez (1999); parameter estimates $\hat{\boldsymbol{\theta}}$ were obtained analytically whenever possible, and otherwise, numerically using a bisection algorithm (for one-parameter models) or adaptive grid search (for multi-parameter models). The exact distribution function of the power-divergence statistic of index λ with power-divergence estimates of index ν was compared to the chi-squared distribution function to which the exact distribution converges asymptotically. Several discrepancy indices were evaluated in the near and far right tails, i.e., in the regions $R_{\mathrm{near}} = (x_{0.90}, x_{0.95}]$ and $R_{\mathrm{far}} = (x_{0.95}, x_{0.99}]$, where $x_{1-\alpha}$ is the value such that $P(\chi_d^2 \leq x_{1-\alpha}) = 1 - \alpha$ and d are the degrees of freedom of the χ^2 distribution. The results were plotted as a function of the parameter estimate $\hat{\boldsymbol{\theta}}$ with which the composite hypothesis was set up.

2   Main Results and Conclusions

We have studied the accuracy of the approximation for each condition: model $\times$ sample size $\times$ statistic index λ $\times$ parameter estimation index ν $\times$ discrepancy criterion. All the results reported here involve average relative errors (AREs). We have analyzed the dependence of the approximation as a function of the estimated parameter $\hat{\boldsymbol{\theta}}$, of the indices λ and ν in the matched ($\lambda = \nu$) and unmatched ($\lambda \neq \nu$) cases, of the sample size, as well as the analysis of the range of $\boldsymbol{\theta}$ for which the asymptotic approximation is accurate. Finally, we have also studied the magnitude of the minimum admissible value for the expected frequency that guarantees an accurate approximation. Our analysis of the small-sample behavior of power-divergence goodness-of-fit statistics with composite hypotheses for a number of MPT models indicated that, despite small variations across models, the asymptotic chi-squared approximation to the exact distribution of the statistic is reasonably accurate (by the criterion that ARE≤ 0.1) provided:

- Parameters are estimated using maximum-likelihood ($\nu = 0$).

- The power-divergence statistic of index $\lambda = 1/2$ is used for assessing significance in the near right tail, or that of index $\lambda = 1/3$ is used for assessing significance in the far right tail.

- The smallest expectation implied by the composite hypothesis exceeds five.

References

Batchelder, W.H. and Riefer, D.M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin and Review*, **6**, 57-86.

García-Pérez, M.A. (1994). Parameter estimation and goodness-of-fit testing in multinomial models. *The British Journal of Mathematical and Statistical Psychology*, **47**, 247-282.

García-Pérez, M.A. (1999). MPROB: Computation of multinomial probabilities. *Behavior Research Methods, Instruments, and Computers*, **31**, 701-705.

García-Pérez, M.A. and Núñez-Antón, V. (2001). Small sample comparisons for power-divergence goodness-of-fit statistics for symmetric and skewed simple null hypotheses. *Journal of Applied Statistics*, **28**, 855-874.

García-Pérez, M.A. and Núñez-Antón, V. (2004). On the exact distribution of goodness-of-fit statistics in multinomial models with composite hypotheses. *The British Journal of Mathematical and Statistical Psychology*. In press.

Larntz, K. (1978). Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. *Journal of the American Statistical Association*, **73**, 253-263.

Read, T.R.C. and Cressie, N.A.C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. New York: Springer-Verlag.

Riefer, D.M. and Batchelder, W.H. (1999). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, **95**, 318-339.

Riefer, D.M. and Batchelder, W.H. (1991). Statistical inference for multinomial processing tree models. In *Mathematical Psychology: Current Developments* (J.-P. Doignon and J.-C. Falmagne, Eds.), 133-335. New York: Springer.
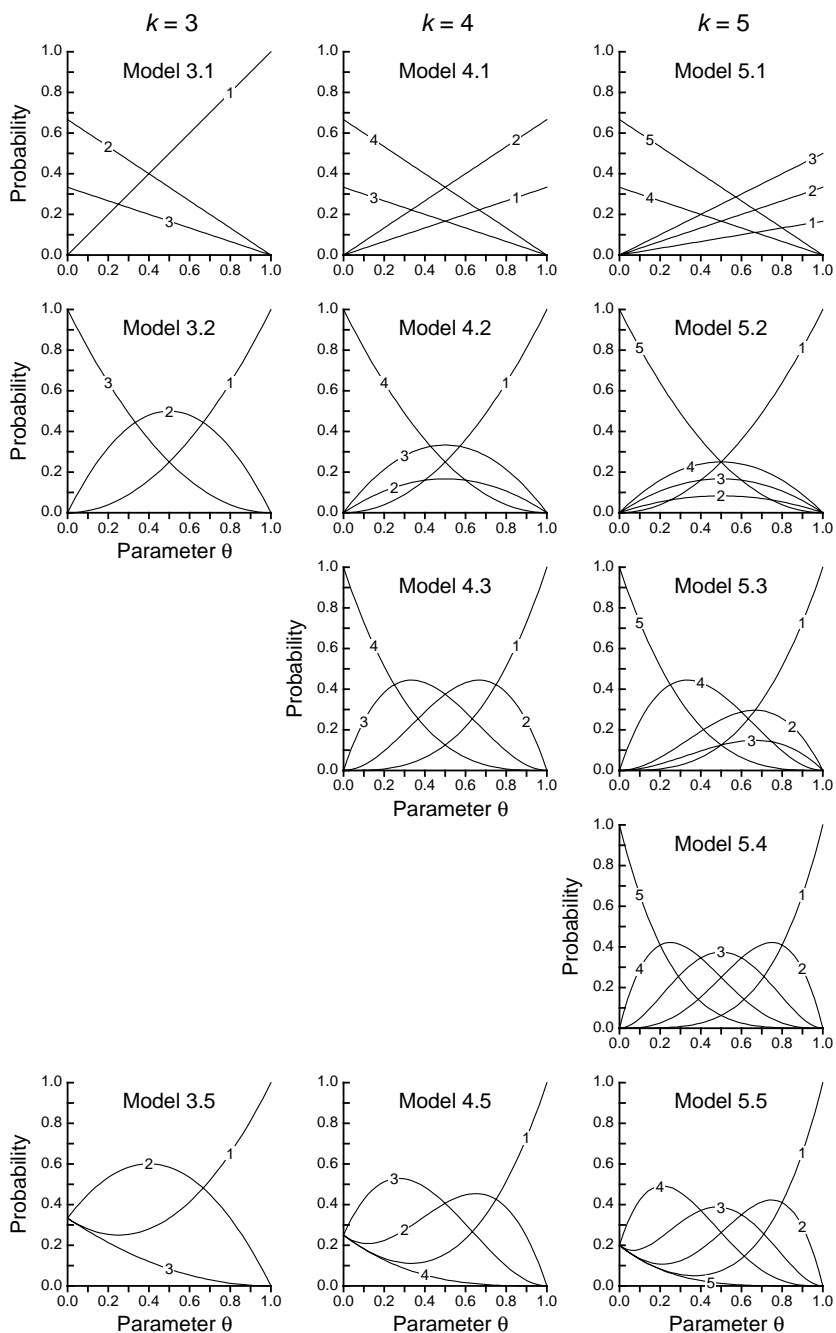
FIGURE 1. Probability distributions as a function of θ in the one parameter-models. Each line pertains to the multinomial cell indicated by the overlaid numeral. Each panel shows a different MPT model. Each column shows all models involving the same number k of cells, with values given at the top. Each row shows models in which cell probabilities are polynomials in θ with the same degree, from first (top row) down to quartic (fourth row). The fifth row shows k-cell models involving polynomials of $(k-1)$-th degree in which the lower boundary of the parameter space renders equiprobability.

A latent variable model of creativity and social compromise

Zoe Georganta[1], Helen Kandilorou[2] and Alexandra Livada[2]

[1] Dept. of Applied Informatics, University of Macedonia, Thessaloniki, Greece
[2] Dept. of Statistics, Athens University of Economics and Business, Athens, Greece

Abstract: This paper develops a latent variable model to investigate the relationship between creativity and social compromise. The theoretical model is then applied to the sophomore population in two large universities, which are located in two major urban centers in Greece. The maximum likelihood estimates of the model are serious indications that young students' creativity may be stifled by a repressive family culture.

Keywords: Latent variables, creativity, learning skills

1 Introduction

'Creativity' is not a 'a flash of inspiration out of the blue' but it relates a concept to a particular body of knowledge, which is as "vital as the novel idea and really creative people spend years and years acquiring and refining their knowledge base - be it music, mathematics, arts, sculpture or design" (Interview for Innovation Exchange, 1999; http://iexchange.London.edu). This is reflected in the now widely accepted definition of innovation equaling creativity plus successful implementation. Creativity cannot be ordered (http://www.eng.uwaterloo.ca/ akay/creative.html notes by Anne K. Gay; http://www.synecticsworld.com/helpdesk/fill-me-in.htm; Jonne Cesevani, 2003, Big Ideas - Putting the Zest into Creativity and Innovation at Work. London, K. Page). It relies heavily on intrinsic motivation (Amabile et al., 1996) and can be stimulated and supported through training and education. Because creativity is an essential building block for innovation, educational systems are committed to encourage its development. However, there are societal characteristics, even in western societies, which tend to stifle creative initiatives especially in young generations. This paper does not aim at developing any new theory of creativity but it seeks to examine the extent to which social compromise in Greece's society affects the creative way of thinking of university students in business, economics and social sciences. The aim of the paper is achieved by developing a latent variable model, which involves the conceptual variables of creativity, social compromise and socio-economic situation. The empirical application of the

model uses the databank DATED (see Databank on Education, 2001 and 2002), which contains 300 variables on motivation, learning skills, psychological and socio-economic factors, score achievements and self-assessment of sophomores in two large Greek universities of business, economics and social sciences. The data is mainly based on two scientific statistical surveys of a control group of 100 students. The ultimate purpose of the surveys is to build a program of learning skill acquisition within the framework of the European Educational Reform, 2002-2006.

2   The Modeling Approach

Latent Variable Modeling (LVM) has been used in social sciences and economics to resolve successfully the problem of statistical and econometric analysis of phenomena, which cannot be accurately expressed in a quantitative dimension only (Georganta, 2003). The LVM approach has been developed mainly by Joreskog and Sorbom (1984), Hayduk (1987) and Bollen (1989), and further discussed and extended by these and other scientists and researchers. LVM uses the analysis of variance-covariance to study the complex path structure of direct and indirect interdependencies of observed factors and their influence on the latent phenomena under investigation. LVM is based on the following three-fold postulation:

1. Formulation of the hypothesis to be investigated as a causal structure among a set of latent variables.

2. Detection of a set of observed factor-variables, which can be used as proxies of the latent variables. Such observed variables are called indicator variables.

3. Specification of the latent variables as functional combinations of the indicator- variables and measurement errors in a causal chain of observed and non-observed variables.

The general form of a latent variable model includes the following three matrix equations:

$$\eta = B\eta + \Gamma\xi + \zeta \qquad \text{structural equation model} \qquad (1)$$

$$y = \Lambda_y\eta + \epsilon \qquad \text{measurement model for } y \qquad (2)$$

$$x = \Lambda_x\xi + \delta \qquad \text{measurement model for } x \qquad (3)$$

where η and ξ are random vectors of latent dependent and independent variables, respectively, B and Γ are coefficient matrices, and ζ is a random vector of disturbance terms. The elements of B represent direct causal effects of η-variables on other η-variables and the elements of Γ represent direct causal effects of ξ-variables on η-variables. The vectors η and ξ are

not observed but instead vectors y and x are observed, such that the two measurement models represented by equations (2) and (3) hold. Λ_y and Λ_x are coefficient matrices, and ϵ and δ are vectors of errors of measurement in y and x, respectively.

The observed vectors y and x contain indicator variables for the unobserved or latent variables η and ξ, respectively. The latent variables correspond to theoretical constructs or variables measured correctly. For this reason, they may be called "true" variables. The structural equation model represented by equation (1) specifies the causal relationship between the "true" or latent variables η and ξ. The measurement models represented by equations (2) and (3) specify how the latent variables, or hypothetical constructs η and ξ, are measured in terms of the observed variables y and x, respectively. It is emphasized that ζ in equation (1) is a vector of classical disturbances, including all random discrepancies that emerge between the actual values of η and the values that would be obtained by the corresponding exact or, in the case of no disturbances, stable functional relationship. Such random discrepancies may be due to omitted variables from the model, or to some "intrinsic" randomness in elements of vector η which cannot be explained anyway, or to any other non-systematic influence on vector η which cannot be captured by the right-hand part of equation (1) no matter how elaborate it is. What ζ does not include is measurement errors, which are instead cast into the vectors ϵ and δ in equations (2) and (3). For the LV model (1)-(3) the following classical assumptions are made:

(a) The error terms ζ, ϵ and δ have zero mean values. ζ is uncorrelated with the vectors ξ and η. ϵ and δ are uncorrelated with the corresponding vectors η and ξ, respectively.

(b) The matrix B has zeroes in the diagonal, and

(c) The matrix $(I - B)$ is non-singular.

Assumptions (a) ensure that equations (1)-(3) are well specified including all the important determinants of the dependent variables. Regarding assumption (b), the elements of matrix B are assumed not to depend on themselves. Assumption (c) is required for estimation purposes, i.e. the inverse of matrix $(I - B)$ or $(I - B)^{-1}$ must exist.

3   The Empirical Model

Following the LVM methodology, as well as Georganta and Hewitt (2004), the following empirical model (4)-(6) is constructed:

$$\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ \beta & 0 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} [\xi] + \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix} \qquad (4)$$

TABLE 1. The notation used

Notation	Description
η_1	Creativity (conceptual variable)
η_2	Social compromise (conceptual variable)
ξ	Socio-economic situation (conceptual variable)
y_1	Index 1 of creativity (constructed by the authors)
y_2	Index 1 of social compromise (constructed by the authors)
y_3	Index 2 of creativity (constructed by the authors)
y_4	Index 2 of social compromise (constructed by the authors)
x_1	Parents education (Index constructed by the authors)
x_2	Parents profession (Index constructed by the authors)

$$
\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \lambda_1 & 0 \\ 0 & 1 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{bmatrix} \tag{5}
$$

$$
\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ \lambda_3 \end{bmatrix} [\xi] + \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} \tag{6}
$$

The notation used is reported in Table 1.

4   The Estimates

The model (4)-(6) is overidentified. It has 21 moments and 16 free parameters to be estimated. These are the six coefficients, β, γ and λ, the variances of the error terms and the variance-covariance matrix of the exogenous indicator variables. The model is estimated by using the software LISREL (www.ssicentral.com/lisrel/mainlis.htm). The maximum likelihood estimates of the model are presented in Table 2.

5   Conclusions

The results in Table 1 show a negative, but statistically significant relationship between creativity and social compromise, implying that Greece's young and educated generations may be suffering a serious stifling of their creativity because of a prevailing repressive attitude within Greek families.

TABLE 2. Maximum likelihood estimates of models (4)-(6)

Parameter	Estimate	T-value
β	-0.617	-3.93
γ_1	0.211	0.54
γ_2	0.550	2.95
λ_1	1.023	2.79
λ_2	0.971	10.32
λ_3	5.163	3.49
χ^2	6.33,	
	degrees of freedom=5	
R^2	0.8865	

References

Amabile, T.M. et al. (1996). Assessing the work environment for creativity. *Academy for Management Journal*, **39**, 5, 1154-1184.

Bollen, K.A. (1989). Structural Equations with Latent Variables. John Wiley & Sons: New York.

DATED Databank on Education (2001, 2002). University of Macedonia of Economic and Social Sciences, (Department of Applied Informatics), Athens University of Economics and Business (Department of Statistics)

Georganta, Z. (2003) . Latent Variable Modeling of Price-change in 295 Manufacturing Industries. *Applied Stochastic Models in Business and Industry*, **19**, 67-88.

Georganta, Z., and Hewitt, W.D. (2004). Information Economy and Educational Opportunities: A Latent Variable Model of Learning Skills (forthcoming). Proceedings Frontiers in Education 2004, IEEE 2004.

Hayduk, L.A. (1987). Structural Equations Modeling with LISREL. John Hopkins University Press: Baltimore.

Joreskog, K.G., and Sorbom, D. (1984). LISREL VI, Analysis of Linear Structural Relationships by the Method of Maximum Likelihood, User's Guide. Scientific Software: Mooresville, IN.

Quasi-likelihood ratio statistic for robust hypothesis testing in the presence of nuisance parameters

Luca Greco and Laura Ventura [1]

[1] Department of Statistics, via C. Battisti 241, 35121 Padova, Italy
(e-mail: greco@stat.unipd.it, ventura@stat.unipd.it).

Abstract: We discuss the problem of robust hypothesis testing about a scalar parameter of interest in the presence of a nuisance parameter. It is well-known that standard likelihood procedures are not robust with respect to model misspecification or the presence of outliers, which can badly affect hypothesis testing and model selection.Therefore, we discuss a quasi-profile loglikelihood with the standard distributional limit behaviour which, at the same time, assures robustness under small departures from the assumed model. This function is based on a profile estimating function, obtained by modifying a generalised profile score. A numerical study and an application about inference on the shape parameter of a gamma model, in the context of modelling personal-income distributions, are also considered.

Keywords: B-robustness; Generalised score function; Likelihood ratio statistic; Profile estimating equation; Pseudo-likelihood.

1   Introduction

Consider a sample $y = (y_1, \ldots, y_n)$ of n independent observations with distribution function $F(y; \theta)$, depending on an unknown parameter $\theta \in \Theta \subseteq \mathbb{R}^p$, $p > 1$. Suppose that θ is partitioned as $\theta = (\tau, \lambda)$, where τ is a scalar parameter of interest and λ a $(p-1)-$dimensional nuisance parameter. A common aim in many studies, such as model selection in nested models, is to check the null hypothesis $H_0 : \tau = \tau_0$ on the parameter of interest. Classical test statistics for this problem are tipically based on a pseudo-likelihood function, i.e. a function of y and τ, having properties similar to those of a likelihood function when there is no nuisance parameter. The most commonly used pseudo-loglikelihood is the profile loglikelihood $\ell_p(\tau) = \ell(\tau, \hat{\lambda}_\tau)$, where $\ell(\theta) = \ell(\tau, \lambda)$ denotes the usual loglikelihood for θ and $\hat{\lambda}_\tau$ is the maximum likelihood estimate (MLE) of λ for fixed τ. Standard likelihood procedures for testing H_0 are then based on the profile likelihood ratio test (LRT)

$$W_p(\tau_0) = 2 \left\{ \ell_p(\hat{\tau}) - \ell_p(\tau_0) \right\} \ , \tag{1}$$

where $\hat{\tau}$ is the MLE of τ. It is well-known that classical inference based on (1) is not robust with respect to model deviations or influential observations. While robust literature offers many solutions for inference on the whole parameter θ (see e.g. Hampel *et al*, 1986), the situation with a nuisance parameter has been somewhat neglected. An exception is given by Heritier and Ronchetti (1994), but their robust version of the LRT does not present the standard asymptotic χ^2 distribution. In view of this, hypothesis testing about τ is often based on Wald-type test statistics. The aim of this contribution is to discuss a robust quasi-likelihood ratio statistic (QLRT) to be used for testing hypothesis about τ, when λ is unknown. The QLRT has a standard χ_1^2 asymptotic distribution and, at the same time, assures robustness under small departures from the assumed model. Since the QLRT discussed in this paper is based on a profile robust estimating function, obtained by modifying a generalised profile score, it can be applied in very general situations ofpractical interest.

2   Background theory

The aim of this section is to derive a robust version of the LRT to be used in hypothesis testing problems, such as model selection in nested models. For example, the interest may lie on the shape parameter when modelling the error distribution of a regression-scale and shape model. Consider a bounded estimating function for θ of the form $\Psi_\theta = (\Psi_\tau(y;\theta), \Psi_\lambda(y;\theta))$. Let $\tilde{\theta}$ be the solution of the unbiased estimating equation $\Psi_\theta = 0$ and let $\tilde{\lambda}_\tau$ be the estimate for λ derived from $\Psi_\lambda = 0$, when τ is considered as known. An estimator $\tilde{\tau}$ for τ with bounded influence function can be obtained as the root of the estimating equation $\Psi_\tau(\tau, \tilde{\lambda}_\tau) = 0$. Such an estimator is called B-robust. A quasi-profile loglikelihood function corresponding to $\Psi_\tau(\tau, \tilde{\lambda}_\tau)$ is (Adimari and Ventura, 2002)

$$\ell_{qp}(\tau) = \int^\tau w(t, \tilde{\lambda}_t)\Psi_\tau(t, \tilde{\lambda}_t)\, dt \; . \tag{2}$$

The scale adjustment $w(\tau, \lambda)$ can be obtained analitically in very simple special cases, but in general we must resort to Monte Carlo simulation (McCullagh and Tibshirani, 1990). In practice, in hypothesis testing problems, it is necessary to obtain a QLRT based on (2) with the classical χ_1^2 asymptotic distribution. In view of this, the QLRT

$$W_{qp}(\tau_0) = 2\{\ell_{qp}(\tilde{\tau}) - \ell_{qp}(\tau_0)\} \tag{3}$$

may be used as an ordinary LRT for testing $H_0 : \tau = \tau_0$, assuring at the same time robustness under small departures from the specified model. A critical region for testing H_0 can be constructed as $\{y : W_{qp}(\tau_0) \geq \chi_{1;1-\alpha}^2\}$, where $\chi_{1;1-\alpha}^2$ is the $(1-\alpha)$−quantile of the χ_1^2 distribution. The main hitch

in using (3) is that, in many problems of practical interest, it can be difficult to find the estimating equations Ψ_τ and Ψ_λ. This is the case, for example, when a shape parameter is of interest. However, this problem can be overcome by a recent approach based on a truncation argument applied to a generalised profile score function (Greco and Ventura, 2004). A generalised profile log-likelihood function $\tilde{\ell}_p(\tau) = \ell(\tau, \tilde{\lambda}_\tau)$ can be obtained by replacing the MLE $\hat{\lambda}_\tau$ with another consistent estimate $\tilde{\lambda}_\tau$ for the nuisance parameter (Severini, 1998). Then a bounded profile estimating function for the interest parameter can be constructed in a standard way by defining an appropriate weighting function $w(\cdot, b)$, which assignes weights in $[0, 1]$ to each component of the generalised profile score $(\partial/\partial\tau)\tilde{\ell}_p(\tau) = \ell_\tau(\tau, \tilde{\lambda}_\tau; y_i)$. The costant $b > 0$ is related to the upper bound imposed on the influence function of $\tilde{\tau}$. The resulting estimating function assumes the form

$$\Psi_\tau(\tau, \tilde{\lambda}\tau) = \sum_{i=1}^{n} w_i(b)\ell_\tau(\tau, \tilde{\lambda}_\tau; y_i) \ . \tag{4}$$

3   Numerical study

Assume that the underlying distribution of the data is a gamma model with unknown parameters, and the shape parameter τ is of interest. To eliminate the scale nuisance parameter λ, we use a MAD-type estimator $\tilde{\lambda}_\tau$, which is Fisher consistent at the gamma model for τ considered as known. The first two plots in Figure 1 show the behaviour of the LRT and of the QLRT under the true model (simulated sample of size $n = 200$) and under a small contamination (replacement of the five larger observations by even larger values). The LRT shifts remarkably, whereas it does not occur for the QLRT. Note that the 0.95-level confidence interval for τ based on the LRT under the contaminated sample does not include the true value of the parameter. The stability of the QLRT can also be assessed by means of an empirical sensitivity analysis. We use a simulated sample of size $n = 100$ from a gamma distribution. The 100th value in the sample is perturbed and allowed to take arbitrarily large values. At each time LRT and QLRT for testing $H_0 : \tau = \tau_0$, where τ_0 is the true parameter value, are recomputed. Last plot in Figure 1 displays the behaviour of the p-value associated to both the LRT and the QLRT. It is evident that the LRT appears sensitive to outlying observations, whereas the p-value associated to the QLRT is more stable. A simulation experiment (based on 3000 Monte Carlo trials) has also been performed in order to evaluate the empirical coverages of the nominal $1 - \alpha$ confidence intervals for the shape parameter obtained by QLRT. The results are given in Table 1 and they indicate that the QLRT performs well both under the true model and under the contaminated model.
For an application to real data, consider the empirical distribution of household incomes in 1979 in UK. We decide to fit a gamma distribution to the
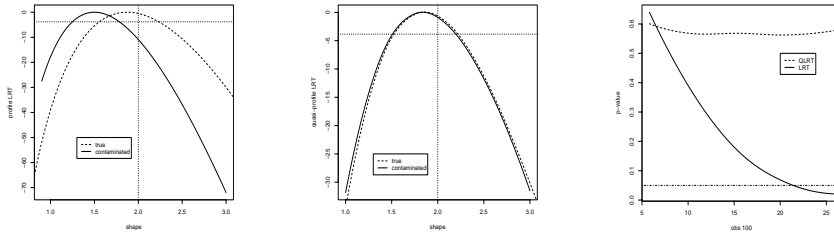
FIGURE 1. LRT (left) and QLRT (middle) for the shape parameter (the horizontal dotted line gives the 0.95 confidence interval); Sensitivity curves (right) of the p-value for LRT and QLRT (the horizontal line corresponds to the 0.05 significance level).

TABLE 1. Empirical coverage probabilities of the confidence intervals for the shape parameter obtained from the QLRT.

distribution	$1 - \alpha$		
	.990	.950	.900
Gamma (2,1)	.991	.956	.909
Gamma (2,1) 3% cont. by Gamma (2, 5)	.989	.947	.893

data (see Victoria-Feser and Ronchetti, 1994). We desire inference on the shape parameter not to be influenced by extreme observations in the tails. Therefore, the weighting function used to bound the generalised profile score functionis choosen so that more importance is given to the most frequent observations, located in the centre of the distribution. In Figure 2 it can benoted that the 0.95-level confidence interval includes the value estimated in Victoria-Feser and Ronchetti(1994). Finally, the plot in Figure 3 gives the histogram of the empirical distribution and the estimated Gamma distribution by our proposal (solid line), the OBRE (dashed line) and MLE (dotted line). The estimated curve according to the MLE tends to be influenced by extreme observations in the tails whereas the distribu-
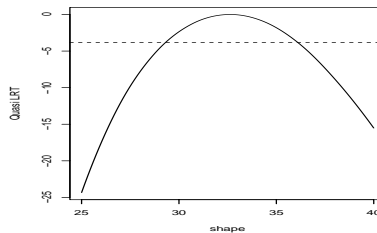
FIGURE 2. QLRT for the shape parameter of the distribution of the household income data (the horizontal dashed line gives the 0.95 confidence interval).
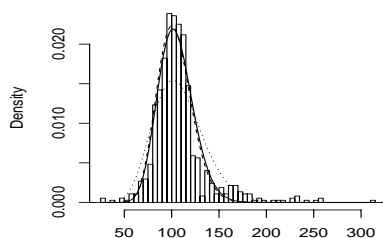
FIGURE 3. Histogram of household income data and estimated Gamma distribution by RGMLE (solid line), the OBRE (dashed line) and the MLE (dotted line).

tions estimated by RGMLE and OBRE catch the inequality structure of the majority of the data.

References

Adimari, G., Ventura, L. (2002). Quasi- profile loglikelihood for unbiased estimating functions, *Ann. Inst. Statist. Math.*, **54**, 235–244.

Greco, L., Ventura, L. (2004).Robustness and estimating equations in the presence of a nuisance parameter,Proceedings of the *XLII Riunione Scientifica della SocietàItaliana di Statistica*, Bari, 9-11 june 2004, to appear.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., Stahel, W. A. (1986). *Robust Statistics. The Approach Based on Influence Functions*, Wiley, New York.

Heritier, S., Ronchetti, E. (1994). Robust bounded-influence tests in general parametric models, *J. Amer. Statist. Assoc.*, **89**, 897–904.

McCullagh, P., Tibshirani, R. (1990).A simple method for the adjustment of profile likelihoods, *J. R. Statist. Soc. B* , **52**, 325–344.

Severini, T.A. (1998). Likelihood functions for inference in the presence of anuisance parameter, *Biometrika*, **85**, 507–522.

Victoria-Feser, M., Ronchetti, E.M. (1994).Robust methods for personal income models, *Can. J. Statist.*, **22**, 247–258.

Microarray Experiments For Gene Expression In Fish Stress Studies

Emma Holian[1] and John Hinde[1]

[1] Department of Mathematics, National University of Ireland, Galway.

Abstract: The aim of this work is to explore various statistical techniques to identify genes which contribute to some change in phenotype level. Experiments are carried out with the aid of microarray technology which allows the simultaneous screening of several thousand of candidate genes. We outline the microarray methodology and how it is applied to the fish stress experiment. To identify which genes display differential expression an ANOVA model is applied to account for some of the systematic variability, such as array or dye effects, these effects being fitted as fixed or random. We also apply multiple testing procedures to address the problems that arise as a result of testing thousands of hypotheses simultaneously.

Keywords: Microarray; Multiple significance testing; Mixed Models.

1 Introduction and Background

The aim of this work is to explore various statistical techniques to identify genes which contribute to some change in phenotype level. This analysis supports an ongoing research project investigating the effects of stress on fish. Samples of different tissues are taken over time from fish kept under controlled conditions. Experiments are carried out using microarray technology to allow the simultaneous screening of several thousands of candidate genes.

A microarray consists of thousands of probes of cDNA, a single stranded copy of genetic material of a known identifiable gene, spotted in an ordered fashion of subgrids on a slide. Hybridization involves the single stranded cRNA of a prepared *target* solution, pipetted onto the slide, binding with its matching single stranded cDNA in the probes to form the double helix DNA molecule. A spot on the slide now gives a measure of the presence and abundance of the genetic material in the target solution. A target solution is made by mixing equal solutions of DNA material from two sources, referred to as the *treatment* and *control*, which are labelled with fluorescent dyes, cyan 3 (green) and cyan 5 (red), to differentiate between them.

After hybridization, the microarray is scanned by a laser, using two frequencies to pick up the signal intensity of each of the two fluorescent dyes, producing *tiff* images. The aim at this point is to compare the intensities

between fluorescent signals at each spot (probe) to assess the level of differential expression, of each gene, in the respective tissue samples which constitute the target solution.

2   Outline of the Fish Stress Experiment

Samples of fish, kept under stressful conditions, were taken at times 2, 6, 24 and 168 hours, tissue material was taken from the brain, pituitary and the liver, separate analyses are carried out for each tissue type. In this experiment we employ a reference design with a dye-swap. For example, at each time-point, the sample from brain tissue is prepared and labelled with red dye. The reference solution is prepared by pooling the samples from all time-points into one sample and labelling with green dye. The first microarray, for the first time-point, is then formed using a sample from that individual time-point (red) and some of the reference solution (green). The procedure is repeated for all time-points producing four microarrays. For the dye-swap part, each slide is repeated with the same tissue samples but with the dyes reversed. This results in eight slides for each tissue type. See Table 1.

TABLE 1. Experiment design for Brain samples of fish stess data.

	Cyan 5 - Red	Cyan 3 - Green
array 1	Time 2 hrs	pooled reference
array 2	Time 6 hrs	pooled reference
array 3	Time 24 hrs	pooled reference
array 4	Time 168 hrs	pooled reference
array 5	pooled reference	Time 2 hrs
array 6	pooled reference	Time 6 hrs
array 7	pooled reference	Time 24 hrs
array 8	pooled reference	Time 168 hrs

3   Array processing - from pixel images to numerical frequencies

The next stage, addressing and segmentation, is an important and difficult phase in analyzing the array. This involves identification of the pixels of the image file which contribute to a spot area against those pixels assigned to background. There are various packages which do this and many different methods, including fixed-circle, histogram method and seeded region growing, (SRG), which is provided within an R-platform package *SPOT*. The output is now in the form of numerical frequency data with two values

for each dye for each spot, a foreground value which is the mean of the frequencies of pixels assigned as spot area and a background value, the mean of the frequencies of pixels assigned as background to that spot.

The data presented at this level is often problematic, mainly due to the large number of genes, the small amount of independent samples and the variability arising at each stage of the process. Examples of this experimental variation include, manual segmentation methods, which may be subject to the experimenter's judgement. Scratches or dirt on the slide distorting the fluorescent signal of affected spots. Uneven washing of the slide, resulting in high background intensities or spatial heterogeneity across the slide. In addition, it is also a known property of the dyes that one naturally gives a higher signal when scanned by the laser. A process known as normalization attempts to reduce non-biological variations in expression, ensuring representation of values on a comparative scale. Possible normalization corrections include background correction, centering methods and scale adjustments. Centering methods are applied, globally between slides, to centre the distribution of logged intensities for each array to zero. Scale methods then adjust for variations in the spread of the logged intensities. These methods can also be applied within a slide, to correct for dye or spatial dependencies or in cases where dye bias depends on intensity strength.

4   Model fitting and testing for differential expression

An ANOVA model is used to identify which genes are displaying differential expression accounting for some of the systematic effects, otherwise amended for at the normalization stage, such as array or dye effects. For example, for the fish-stress data, to account for array, dye, variety and time effects, $Time_{tg}$, we fit the gene-specific model

$$log_2(y_{ijkgt}) = G_g + AG_{ig} + DG_{jg} + VG_{kg} + Time_{tg} + \epsilon_{ijkgt} \qquad (1)$$

In this model G_g is an overall mean for logged frequencies, $\{y_{...g.}\}$, for gene g, AG_{ig} are gene-specific array effects for arrays i, DG_{jg} are gene-specific dye effects for dyes $j = 1, 2$, and VG_{kg} are gene-specific variety effects for varieties $k = 1, 2$. It is this term that is of interest as the resulting values estimate the gene expressions for *treatment* and *control* (or *reference* in this case) and so can estimate the magnitude of differential expression $VG_{1g} - VG_{2g}$ for a gene g. In order to estimate these parameters we apply the following constraints

$$G_g = \sum_{i=1}^{8} AG_{ig} = \sum_{j=1}^{2} DG_{jg} = \sum_{k=1}^{2} VG_{kg} = \sum_{t=1}^{4} Time_{tg}. \qquad (2)$$

Note that arrays are nested within time so we interpret the term AG_{ig} as array effects within time.

We also extend the model by fitting AG_{ig} as a random effect. Using prior estimates for the random effect and error variances obtained by fitting the full fixed model and solving Henderson's equations, we obtain the BLUE and BLUP estimates of the fixed and array effects respectively. These estimates are then updated by iterative procedures to maximize the log likelihood or the restricted log likelihood functions.

To test for differential expression the analysis also produces three forms of F-statistic. These each have a different allowance for knowledge drawn from testing all genes simultaneously, that is, they use different weighted combinations of the gene-specific variance and global variance estimates. Fitting a full fixed effects model, the distributions of these F-statistics are estimated by random permutations of the labels *Treatment*s and *Control*s for the frequency data. While fitting the random effects model these distributions are assumed to be known distributions. A volcano plot displays the p-values of all three statistics simultaneously for all genes.

5   Multiple testing problems

To allow for multiple significance testing we use two procedures, Westfall and Young step-down permutation procedure and another technique known as Significance Analysis of Microarrays (SAM). Both of these procedures are implemented in the R system as part of the Bioconductor package. These methods address the problems that arise as a result of testing thousands of hypotheses simultaneously and attempt to apply some control of the number of Type I errors that may occur. In particular, SAM adjusts the individual t-statistics for differential expression of each gene using information obtained globally across all genes, shrinking the test statistic for genes where the estimated standard deviation is close to zero. Under the null hypothesis of no differential expression, the distribution of the t-statistic is calculated, for each gene, by permutations of sample labels. Significance cut-offs are calculated while controlling the positive false discovery rate, pFDR. This is a measure of the number of genes falsely called significant as a proportion of the total number of genes called significant.

6   Remarks

An interesting aspect of the model is its potential to offer insight into the expression patterns of genes over time, not only to classify genes by similarities in expression patterns, but also to model these patterns as specified functions.

The aspects outlined above are the preliminary investigations of ongoing research. Full results and conclusions from comparisons between methods will be presented.

Although the results relate to the fish stress data set, we are also looking at data from another interesting application of microarrays. The aim here is to investigate gene expression levels in reproductive tissues during pregnancy and labour, as part of a study of disorders associated with pregnancy, such as premature labour and pre-eclampsia. Data in this experiment comes from two groups of patients, pregnant labouring and pregnant non-labouring women. Endometrial tissue samples were extracted, from individuals in the pregnant-labouring, treatment group, during emergency caesarean section where labour had started naturally. Similarly, for individuals in the pregnant non-labouring, control group, endometrial tissue samples were taken during a scheduled caesarean section where the labour process had not started. This is an ongoing experiment and very limited data are currently available.

References

Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects (with discussion). *Statistical Science*, **6**, 15-51.

Dudoit, S., Luu P., Yang, Y.H., Speed, T.P. (2001). Normalization for cDNA microarray Data *Microarrays: Optical Technologies and Informatics*, 4266.

Ge, Y., Dudoit, S., Speed, T.P.](2003). Resampling-based Multiple Testing for Microarray Data Analysis *Test*, **12**, 1-77.

Gentleman, Rossini, Dudoit, and Hornik (2003). The Bioconductor FAQ *http://www.bioconductor.org*

Kerr, M.K., Churchill, G.A., Afshari, C.A., Bennett, L., et.al (2000). Statistical Analysis of a Gene Expresssion Microarray Experiment with Replication *Statistica Sinica*, **12**, 203.

Storey, J.D., and Tibshirani, R. (2003). Statisical sigificance for genomewide studies *Proc. Natl. Acad. Sci.*, **100**, 9440-9445.

Tusher, V.G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response *Proc. Natl. Acad. Sci.*, **98**, 5116-5121.

Wolfinger, R.D., Gibson, G., and Wolfinger, E.D. (2001). Assessing Gene Sinificance from cDNA Microarray Expression Data via Mixed Models *Journal of Computational Biology*, **8**, 625-637.

A growth mixture model for multivariate outcomes : application to cognitive ageing

Cécile Proust[1] and Hélène Jacqmin-Gadda[1]

[1] INSERM E0338, ISPED, 146 rue Léo Saignat, 33076 Bordeaux cedex, France.

Abstract: In this work we propose a heterogeneous linear mixed model for multivariate longitudinal data using a latent process in order to describe the evolution of cognitive functions in a cohort of initially non-demented elderly people.

The latent process, which represents the unobserved global cognitive level, is defined by random effects whose distribution is a mixture of Gaussians. The unobserved global cognitive level is assessed using a battery of psychometric tests, each test representing a distinct measure of the global cognitive level.

The joint modelling proposed in this work allows us to exploit information contained in several psychometric tests in order to estimate distinct profiles of the global cognitive evolution. The mixture of distributions also allows us to classify the subjects according to these profiles and to characterize their evolution.

The growth mixture model using the Mini Mental State Examination and the Isaacs Set Test highlights two distinct courses of the global cognitive level. The first profile has a slight decline and the second a sharp decline until the last visit. This model gives a very clear classification and the subjects classified in the second class have a higher risk of dementia, death or disablement.

Keywords: mixture model; random effects; joint modelling; classification; dementia

1 Introduction

Cognitive ageing is a continuous process which has to be studied with longitudinal methods in order to take into account the variability of the evolutions between the subjects. To achieve this, mixed models (Laird and Ware, 1982) have been widely used. However besides this variability, there exists an extra heterogeneity in the population due in particular to the presence of people with pathological and normal cognitive ageing. To take into account this heterogeneity, mixed models with a mixture of distributions for the random effects can be used (Verbeke and Lesaffre, 1996; Muthén and Shedden, 1999). This kind of model enables us not only to estimate distinct curves in the population but also to classify subjects from those curves.

In epidemiological studies, cognitive ageing is assessed using psychometric tests. These tests are different measures of the global cognitive level, which itself is not observed.

The aim of this work is to propose a heteregeneous linear mixed model for multivariate longitudinal data using a latent process in order to describe distinct profiles of evolution for the global cognitive level. The model is applied to data from a cohort of initially non-demented elderly subjects by using repeated measurements of several psychometric tests. The latent process is defined by random effects whose distribution is a mixture of Gaussians and the different psychometric tests are linear transformations of the latent process, measured with error.

2   Model

Let $\Lambda_i(t)$ be the latent process which represents the unobserved trajectory of global cognition for the subject i, $i = 1, ..., N$ and t is the time. The growth mixture model or heterogeneous mixed model is defined as :

$$\Lambda_i(t) = u_{0i} + u_{1i}t + u_{2i}t^2 + X_i(t)\beta \tag{1}$$

$X_i(t)$ is the p-vector of covariates associated with the vector of fixed effects β. The distribution of the vector $u_i = (u_{0i}, u_{1i}, u_{2i})^t$ of random effects is a mixture of G multivariate Gaussians with means $(\mu_g)_{g=1,...,G}$ and a specific covariance matrix $\omega_g D$, where $(\omega_g)_{g=1,...,G}$ are scalars. Thus

$$u_i \sim \sum_{g=1}^{G} \pi_g N(\mu_g, \omega_g D) \tag{2}$$

with $\omega_1 = 1$ so the matrix D is the covariance matrix for the first component. D is unstructured except that the variance of the random intercept for the first component is constrained to 1. The vector $(\mu_{0g})_{g=1,...,G}$ satisfies the condition $\sum_{g=1}^{G} \mu_{0g} = 0$. Each component g of the mixture has a probability π_g with $0 \leq \pi_g \leq 1$, $\forall g = 1, ..., G$ and $\sum_{g=1}^{G} \pi_g = 1$.

Let $Y_i^k = (Y_{i1}^k, ..., Y_{in_i^k}^k)$ be the response vector of the n_i^k measurements of the subject i for test k, $k = 1, ..., K$. Then, we assume

$$Y_{ij}^k = J_k + L_k \Lambda_i(t_{ij}^k) + e_{ij}^k \tag{3}$$

where J_k is an intercept and L_k a scale parameter for test k ; $t_i^k = (t_{i1}^k, ..., t_{in_i^k}^k)$ is the n_i^k-vector of measurement times for test k. The errors e_{ij}^k are assumed to be independently normally distributed with mean zero and variance σ_k^2.

3   Estimation

The estimation of the model is performed with a fixed number of components G. The parameters are estimated using the maximum likelihood

method. The observed log-likelihood, which has a closed form since the marginal distribution of Y_i^k is a mixture of multivariate Gaussians, is maximized directly using an improved Marquardt algorithm developed in Fortran90. A Marquardt algorithm is a Newton-Raphson like algorithm in which the Hessian is inflated if necessary to make it positive definite when updating the parameters. We added a linesearch for the step to ensure that the likelihood increased at each iteration.

A logistic transformation of $(\pi_g)_{g=1,\ldots,G-1}$ ensures that the probabilities are between 0 and 1 and the Cholesky transformation of D ensures the positivity of the covariance matrix.

Posterior individual probabilities $\hat{\pi}_{ig}$ are computed using Bayes Theorem from the data and the estimated parameters (see Verbeke and Lesaffre, 1996). Then, the subjects are classified into profiles according to the largest posterior probability.

4   Application

The objective of the application is to describe the distinct profiles of evolution of the global cognitive level in a cohort of non demented elderly people. The classification of subjects given by the mixture model is also compared with the dementia diagnosis at the end of the follow-up, to assess if this method can be a predictive tool of dementia diagnosis.

Data come from the French prospective cohort study PAQUID initiated in 1988 to study normal and pathological ageing (see Letenneur *et al*, 1994). Subjects were interviewed at beaseline and were seen again 1 (T1), 3 (T3), 5 (T5), 8 (T8) and 10 (T10) years later. Two psychometric tests are considered : the Mini Mental State Examination (MMSE), which evaluates the global cognitive performance, and the Isaacs Set Test (IST), which is a test of verbal fluency. The subjects included in this study have a negative dementia diagnosis at the visit T5 and have a diagnosis of dementia at the visit T8. They also have at least one measurement at each test during the follow-up of 7 years (between T1 and T8). This leads to a sample of 1382 subjects having between 1 and 4 measures per test. The time is defined as the negative time between the measurement and the last visit T8.

The model contains a linear function of time, an effect of educational level, occupation and age (older or younger than 80 years old at the the last visit) and an age-time interaction.

The growth mixture model with two components of mixture was fitted and the Bayesian Information Criterion (BIC) was substantially better than for the homogeneous mixed model ($\Delta BIC = 52.8$). Two distinct courses of the global cognitive function were clearly distinguished, with class probabilities of 0.96 and 0.04. Figure 1 represents the estimated mean curves for the two profiles for each psychometric test. The cognitive tests for the first profile slightly decrease until the last visit, whereas for the second profile, the
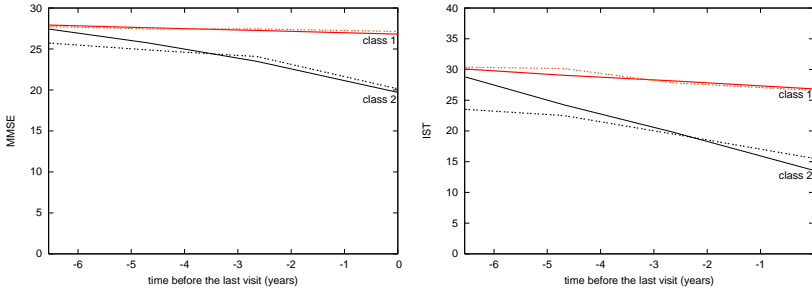
FIGURE 1. Posterior-probability-weighted sample means (dashed line) and estimated mean curves (plain line) of the two components (class 1 and class 2) for MMSE (left) and IST (right).

TABLE 1. Classification table for the growth mixture model with two components.

	mean probability to belong to:	
	class 1	class 2
subjects in class 1	**0.987**	0.013
subjects in class 2	0.074	**0.926**

cognitive tests which are lower 7 years before sharply decrease until the last visit.

An assessment of the classification was performed using some of the methods described in Muthén *et al*, 2002. For subjects classified in class 1 and class 2, Table 1 presents the averages of the posterior probabilities to belong to each class. It reveals very high diagonal values which indicates a good classification quality. Then, the entropy measure defined in (4) is also very high ($E_2 = 0.94$) which indicates a clear discrimination.

$$E_G = 1 - \frac{\sum_i \sum_g -\hat{\pi}_{ig} ln(\hat{\pi}_{ig})}{n ln(G)} \tag{4}$$

The estimated mean curves compared with the posterior-probability- weighted sample mean at each visit (see figure 1) shows that the model fit well the data.

Among the 1,382 subjects, 130 (10.4%) were classified in the second component with the sharp decline. Assuming that this component represents the pathological decline to dementia, we compared the classification with the positive dementia diagnosis at the end of the follow-up to assess if the model was a predictive tool of dementia. The results are in Table 2. The sensitivity of the classification is quite good (65%), the specificity is high

TABLE 2. Relationship between the classification stemmed from the growth mixture model and the dementia diagnosis at the end of the follow-up.

| | dementia diagnosis | | |
classification	positive	negative	total
class 1	23	1229	1252
class 2	43	87	130
total	66	1316	1382

(98%) but the predictive positive value is poor (33%).

5   Conclusion

In this paper, using a growth mixture model and the information contained in two psychometric tests, we described the different profiles of the unobserved global cognitive level in a cohort of initially non-demented people. Two distinct profiles were distinguished : first, a slight decline until the last measurement and secondly, a sharp decline until the last measurement. The discrimination, as assessed by various approaches, was very good.
The second profile could be interpreted as a pathological decline to dementia. But the comparison of the classification with the diagnosis at the last visit shows that it does not highlight directly the subjects who have a positive dementia diagnosis at the end of the follow-up, but a more general pathological cognitive ageing : those people have a higher risk of dementia in the three years after the end of the follow-up, have a higher risk of disablement and have a higher risk of death.

References

Laird, N.M, Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963-74.

Letenneur, L., Commenges, D., Dartigues, J.-F. and Barberger-Gateau, P. (1994). Incidence of dementia and Alzheimer's disease in elderly community residents of south-western France. *Int J Epid*, **23**, 1256-61.

Muthén, B., Shedden, K. (1999). Finite mixture modeling with mixture outcomes using a EM algorithm. *Biometrics*, **55**, 463-9.

Muthén, B., Brown, C.H., Masyn, K., Jo, B., *et al.* (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics*, **3**, 459-75.

Verbeke, G., Lesaffre, E. (1996). A linear mixed-model with heterogeneity in the random-effects population. *JASA*, **91**, 217-21.

Exact Bayesian Inference for Bivariate Poisson data

Dimitris Karlis [1] and Panagiotis Tsiamyrtzis[1]

[1] Department of Statistics, Athens University of Economics and Business, 76, Patission Str., 10434, Athens, Greece

Abstract: We propose Bayesian inference for bivariate Poisson models that generalizes the existing approaches in two important directions. Firstly we propose exact inference contrary to the MCMC approaches existing in the literature and secondly we use a prior distribution that allows for dependencies among the parameters of interest. Our prior is in fact a mixture of priors and the resulting posterior generalizes the idea of conjugacy in the sense that it is again a mixture of the same family but with more components. Computational details and a real data illustration are provided. Extensions of our approach to certain other models is discussed.

Keywords: multivariate gamma distribution; count data;

1 Introduction

The random variables X, Y follow jointly a bivariate Poisson distribution if their joint probability function is given by

$$P(X = x, Y = y) = e^{-(\theta_1+\theta_2+\theta_3)} \frac{\theta_1^x}{x!} \frac{\theta_2^y}{y!} \sum_{k=0}^{min(x,y)} \binom{x}{k} \binom{y}{k} k! \left(\frac{\theta_3}{\theta_1\theta_2}\right)^k.$$

where $\theta_i > 0$, $x, y = 0, 1, \ldots$, denoted as $BP(\theta_1, \theta_2, \theta_3)$. If $\theta_3 = 0$ then the two variables are independent. For a comprehensive treatment of the bivariate Poisson distribution and its multivariate extensions the reader can refer to Kocherlakota and Kocherlakota (1992). Inference for the bivariate Poisson model is not an easy task. The sum appearing in the probability function, the likelihood function is very complicated and in fact it involves n summations, where n is the sample size. To avoid this difficulty, a data augmentation scheme based on the trivariate reduction derivations of the distributions has been considered, for both ML (Karlis, 2003) through an EM algoritm and Bayesian inference (Tsionas, 1999) through an MCMC approach. While MCMC offers some advantages, it can have bad mixing properties, since if the correlation is not large the chain can be trapped, and in this case a large number of iteration may be needed to ensure convergence. The aim of the present paper is to provide relatively easy exact Bayesian inference for the bivariate Poisson model with $\theta_3 > 0$.

2   Likelihood

We will rewrite the likelihood using recursive relationships for deriving the coefficients of the polynomial involved. Namely we prove the following Lemma.

Lemma: Define $v_r^{(n)} = \frac{1}{(x_n-r)!(y_n-r)!r!}$. Given a random sample of size n the likelihood can be written in the form

$$L_n(\theta, \mathbf{X}) = exp\left(-n(\theta_1 + \theta_2 + \theta_3)\right) \theta_1^{\Sigma x_i} \theta_2^{\Sigma y_i} \sum_{k=0}^{S} w_k^{(n)} \left(\frac{\theta_3}{\theta_1\theta_2}\right)^k,$$

where $S = \sum_{i=1}^{n} min(x_i, y_i)$ and $w_k^{(n)}$ are coefficients that can be obtained recursively using

$$w_k^{(n)} = \sum_{r=max\{0,k-s_n^*\}}^{min\{k,s_n^*\}} v_r^{(n)} w_{k-r}^{(n-1)}$$

where $s_i = min\{x_i, y_i\}$, $S_k = \sum_{i=1}^{k} s_i$, $s_n^* = min\{s_n, S_{n-1}\}$ and $w_k^{(1)} = v_k^{(1)}$

3   Bayesian Modelling

Assume the likelihood of $(X,Y)|(\theta_1, \theta_2, \theta_3)$ given in (2). Then, assume that the joint prior for θ_i's $i = 1, 2, 3$ has joint density

$$
\begin{aligned}
\pi(\theta_1, \theta_2, \theta_3) &= \sum_{j=0}^{r} w_j \left(\theta_1^{\alpha_1-j-1} \exp\{-\theta_1\beta_1\}\right) \left(\theta_2^{\alpha_2-j-1} \exp\{-\theta_2\beta_2\}\right) \\
&\times \left(\theta_3^{\alpha_3+j-1} \exp\{-\theta_3\beta_3\}\right),
\end{aligned}
$$

where $\alpha_1 > r$, $\alpha_2 > r$, $\alpha_3 > 0$, $\beta_i > 0$, $i = 1, 2, 3$, $p_j \geq 0$, $j = 0, \ldots, r$, $\sum_{j=0}^{r} p_j = 1$ and

$$w_j = p_j \frac{\beta_1^{\alpha_1-j}\beta_2^{\alpha_2-j}\beta_3^{\alpha_3-j}}{\Gamma(\alpha_1 - j)\Gamma(\alpha_2 - j)\Gamma(\alpha_3 + j)} \qquad for \quad j = 0, 1, \ldots, r.$$

Clearly r determines the number of components in the prior. Then, the posterior distribution will have the form

$$p(\theta_1, \theta_2, \theta_3 | (x, y)) = \sum_{k=0}^{s+r} \rho_k G(\alpha_1+x-k, \beta_1+1) G(\alpha_2+y-k, \beta_2+1) G(\alpha_3+k, \beta_3+1)$$

where

$$\rho_k^* = \left[\sum_{l=max\{0,k-s^*\}}^{min\{k,s^*\}} v_l w_{k-l}\right] \Gamma(\alpha_1+x-k)\Gamma(\alpha_2+y-k)\Gamma(\alpha_3+k) \left(\frac{(\beta_1+1)(\beta_2+1)}{\beta_3+1}\right)^k$$

and $\rho_k = \frac{\rho_k^*}{\sum_{l=0}^{s+r} \rho_l^*}$,   for $k = 0, 1, \ldots, s+r$, $s^* = min(r, s)$.

It is interesting to point out that the prior is a finite mixture of condition-ally independent Gamma densities. The joint prior can be correlated since the mixing operation introduces covariance between the θ's. Assuming a degenerate mixing distribution (i.e. $r = 0$) we obtain that the parameters are independent. The form of the prior can provide a flexible multivari-ate family of gamma distributions with certain desirable properties for real application, like multimodality, variety of shapes, positive and negative correlation etc. Details can be found in a forthcoming article. It is also in-teresting that the posterior density is again a finite mixture of conditionally independent Gamma densities, though now the number of components has changed. The moments of the posterior density can be easily derived via conditioning arguments. The proposed distribution generalizes the idea of non-central gamma densities to more dimensions.

Computationally, one can proceed recursively, by updating the posterior adding one data point at time. This is totally equivalent to using the like-lihood defined in section 2 via recursion. An interesting result is that the number of components in the posterior depends on the data and precisely equals $\sum \min\{x_i, y_i\} + r + 1$, where $r + 1$ is the number of components of the prior.

4   Application

The data refer to the demand for Health Care in Australia, taken by Cameron and Trivedi (1998). We will use two variables, namely the num-ber of consultations with a doctor or a specialist and the total number of prescribed and non-prescribed medications used in past 2 days ($n = 5190$). It is interesting that the data are correlated, the Pearson correlation co-efficient being equal to 0.27 indicating moderate correlation. A bivariate Poisson model is plausible due to the correlation. We applied the exact Bayesian approach discussed in previous section. As priors we used two different sets of independent gamma priors $Gamma(a_i, b_i)$ for each param-eter θ_j, $j = 1, 2, 3$, with hyperparameters $a_i = b_i = 1, 10$ respectively, for $i = 1, 2$. The second set of hyperparameters is more informative in the sense that the prior variance is small.

According to the findings of the previous section, the posterior distribution is a finite mixture with 1076 components for both priors. For the priors, one can easily verify that $\sum_{i=1}^{5190} min(x_i, y_i) = 1075$. The marginal posteriors are respectively ($a_i = b_i = 1, 10$):

$$f(\theta_1) = \sum_{k=0}^{1075} \pi(k) Gamma(a_1 + 1566 - k, b_1 + 5190)$$

TABLE 1. Posterior summaries for the health data using two different priors

	mean	variance	5 % per.	95% per.
$a_1 = a_2 = a_3 = b_1 = b_2 = b_3 = 1$				
k	649.29	619.15	602	693
θ_1	0.1767	$5.7033\ 10^{-5}$	0.1645	0.1893
θ_2	1.0931	$2.3356\ 10^{-4}$	1.0682	1.11845
θ_3	0.12527	$4.7109\ 10^{-5}$	0.11411	0.13671
$a_1 = a_2 = a_3 = b_1 = b_2 = b_3 = 10$				
k	651.82	603.067	603	692
θ_1	0.1772	$5.6483\ 10^{-5}$	0.1655	0.1902
θ_2	1.0925	$2.3240\ 10^{-4}$	1.06765	1.117755
θ_3	0.1272	$4.6778\ 10^{-5}$	0.1162	0.1387

$$f(\theta_2) \ = \ \sum_{k=0}^{1075} \pi(k)Gamma(a_2 + 6323 - k, b_2 + 5190)$$

$$f(\theta_3) \ = \ \sum_{k=0}^{1075} \pi(k)Gamma(a_3 + k, b_3 + 5190)$$

Summary statistics of the posterior densities can be read in Table 1. There are only slight differences between the two different priors, mainly because of the large sample size. Plots of the marginal posteriors can be seen in Figure 1 for both sets of hyperparameters. The posteriors differ slightly mainly because the second set of hyperparameters was very informative. The upper left plot shows the probability function of k shifted by 400 to the left.

5   Discussion

The idea described in the present paper is mainly that of using mixtures of conditionally independent conjugate densities for exact Bayesian inference for the bivariate Poisson model. To this extend the idea of mixture of conjugate priors of Dalal and Hall (1983) is generalized. However, the ideas discussed in the present paper can be extended beyond this model towards certain directions as for example other models with a sum in their likelihood, such as mixture models.

Our procedure is exact and does not rely on MCMC. Computationally is quite easy using the recursions discussed in the paper. Of course MCMC offers the ability to estimate certain other measures of interest but for the specific model it may be trapped and become very slow. We would like also to mention that our approach differs from others in the fact that we start from the full bivariate model with $\theta_3 > 0$ instead of starting from
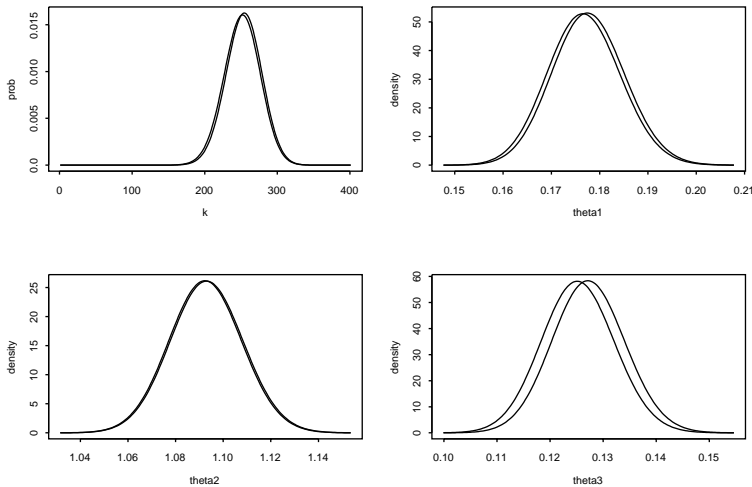
FIGURE 1. Posterior densities for the parameters and the weighting function. The density in the upper left figure is in fact π_k, the mixing distribution of the resulting gamma mixture

the independent Poisson model ($\theta_3 = 0$) and modelling the correlation of the two variables through a common mixing distribution as is usually done (e.g. Chib and Winkellman, 2001).

References

Cameron, A.C. and Trivedi, P.K. (1998). *Regression analysis of count data.* Oxford University Press.

Chib, S and Winkelmann, R. (2001). Markov Chain Monte Carlo Analysis of Correlated Count data. *Journal of Business and Economic Statistics*, **19**, 428-435.

Dalal, S.R. and Hall, W.J. (1983). Approximating priors by mixture of natural conjugate priors. *Journal of the Royal Statistical Society*, **B 45**, 278–286

Karlis, D. (2003). An EM Algorithm for Multivariate Poisson Distribution and Related Models. *Journal of Applied Statistics*, **30**, 63–77.

Kocherlakota, S. and Kocherlakota, K. (1992). *Bivariate Discrete Distributions.* Marcel Decker, NY.

Tsionas, E.G. (1999). Bayesian Analysis of the Multivariate Poisson Distribution. *Communic. in Statistics - Theory and Methods*, **28**, 431–451.

An evaluation of classification techniques applied to the field of NIR/IR spectroscopy

Martin Kidd[1]

[1] Centre for Statistical Consultation, University of Stellenbosch, Private Bax X1, Matieland 7602, South Africa. e-mail: mkidd@sun.ac.za

Abstract: This paper compares various classification techniques applied to data from the field of NIR spectroscopy. It is shown that techniques like MARS and SVM perform better than SIMCA (currently the most popular technique) on 2 sets of simulated data and one set of real data from the wine industry.

1 Introduction

Near infrared (NIR) spectroscopy instruments are used as a non-destructive method for predicting various characteristics of foodstuffs. The data sets used for calibrating these instruments consist of absorption values at different wavelengths (predictor variables) and one or more corresponding measured characteristics (target variables). The target variable can be either a continuous (regression) or categorical (classification) variable. In this paper we focus on the classification problem.

The problem that arises with the data is that of multicollinearity. In chemometrics the method of Simple Modelling of Class Analogy (SIMCA) has become the standard for calibrating NIR instruments on classification problems. Comparative studies have been done in the past to compare various techniques with one another in the NIR classification role. Techniques that were compared were SIMCA, linear discriminant analysis, neural networks, and K-nearest neighbours.

In recent times other techniques have been cited as good classification techniques. These include support vector machines (SVM), boosting and additive trees, and multivariate adaptive regression splines (MARS). In this study, SIMCA is compared with the above mentioned techniques in terms of classification ability. The techniques included in the study were SIMCA, MARS, SVM, multiple additive regression trees (MART), classification trees (CART) and neural networks.
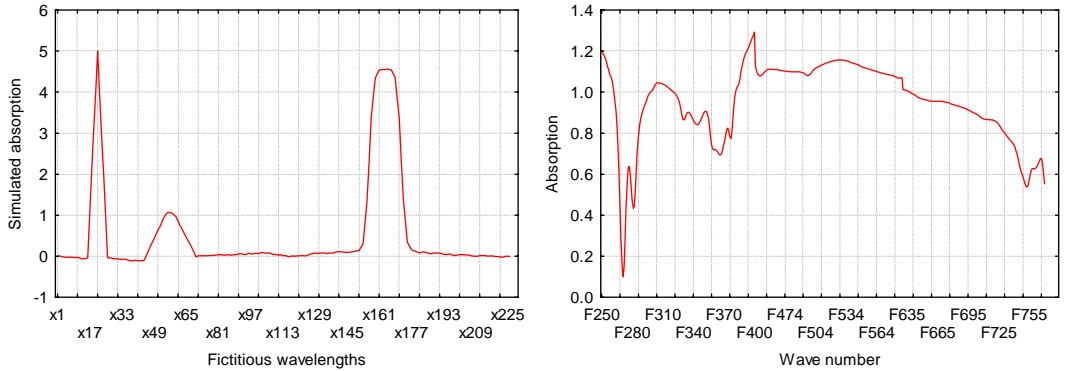
FIGURE 1. Example of simulated data (left panel) and wine data (right panel)

2 Data sets used for comparison

For comparison purpose, simulated data was used to compare the techniques. Figure 1 (left panel) shows an example of one of the simulated cases. A binary classification problem was simulated. Differences in absorption were simulated in 2 different areas of the wave band. The first was at around wave number 17 where a sharp peak in the absorption was simulated. The second was around wave number 161 where a more gradual peak was simulated. One data set with real data was also included in the study. This data set contained wine samples, some which were wood matured, and others which were not wood matured. The right panel in figure 1 shows an example of one of the wine samples. The data set consisted of 54 wood matured samples and 28 non wood matured samples.

3 Method of comparison

The data was randomly divided into a training- and test set (80/20% for simulated data, 50/50% for wine data). Calibration models were derived for each of the techniques from the training set and applied to the test set. The proportions correctly classified for each of the 2 classes (p_1 and p_2) were calculated from the test set. It is important for a good classification technique to have high values for p_1 and p_2. For that purpose an adapted accuracy measure was used which places a penalty on differences between p_1 and p_2. It is defined as: Adapted accuracy $= (p_1 + p_2)/2 - \mathrm{abs}(p_1 - p_2)$.

The above process was repeated n times resulting in n values for p_1 and p_2. Bootstrap averages and confidence intervals were then calculated on the n repetitions for comparison purposes.

4   Modelling techniques included in the study

The following techniques were included in the comparative study:

SIMCA: SIMCA uses principal component analysis (PCA) as basis for its classification model. Different PCA models are fitted for each of the 2 subsets defined by the two classes. A new case is classified by calculating a distance measure of the new case to each of the PCA models. The case is then classified as belonging to the class with the minimum distance.

MARS: MARS is an extension of piecewise linear regression. In piecewise linear regression, more than one regression line is fitted to the data to account for non-linear relationships. The position where one regression line stops and the next line starts, is called a knot position. In the traditional piecewise regression setting, the knot positions must be chosen beforehand. MARS on the other hand, derives the knot positions from the data. MARS can also handle more than one predictor variable as well as combinations of categorical and continuous predictors. In the binary classification setting, the 2 classes of the dependent variable is coded as 0 and 1. A threshold value can then be selected to classify a new case. In this study a threshold value of 0.5 was always used.

CART: CART follows a strategy of repeated binary splits of the data based on optimally selected predictor variables and split values for each variable. When the data is split into 2 sections, the split is made such that the proportion of cases belonging to class 1 is maximised in one section, and vice versa for the other section. The splitting is repeated until some stopping criteria is satisfied, and in this process a binary tree is built based on the data. This tree is then subsequently used to classify new cases.

MART: MART uses the principle of boosting where the purpose is to sequentially apply a classifier to repeatedly modified versions of the data. This sequence then forms a committee of classifiers where the predictions of all of them are combined in a weighted majority vote for the final classification. The modifications to the data are done by assigning weights to each data points in such a way that points that were classified incorrectly by the previous classifier in the sequence, have their weights increased, and points that were classified correctly have their weights decreased. Specifically, in MART, regression trees are sequentially applied to the residuals of the previous tree (called gradient boosting trees) to build the model. Although this method in principle applies to the regression case, it was extended to handle classification problems as well.

Neural networks: Neural networks attempts to emulate the human brain through a network of weights and transfer functions. The network consist
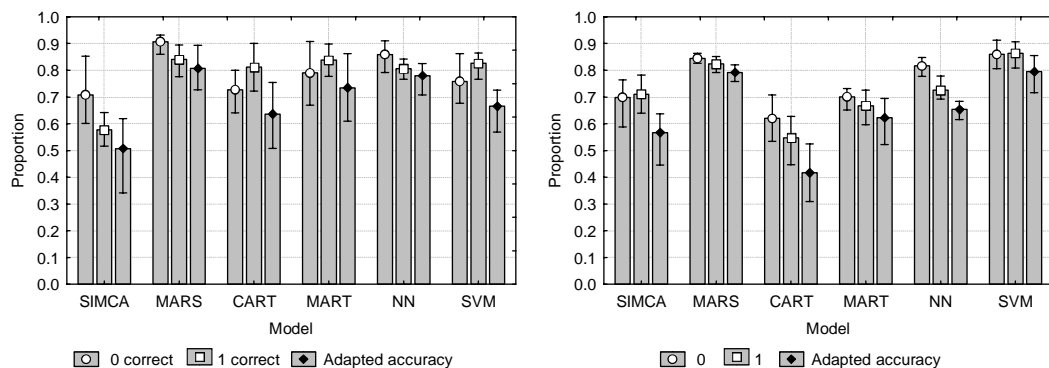
FIGURE 2. Results for the 2 simulated data sets. The left panel shows results for case where absorptions differed at wave number 17, and the right panel for wave number 161.

of an input layer of nodes, one for each predictor, a hidden layer and an output layer of 2 nodes, one for each of the 2 classes. Each of the nodes consists of weights and transfer functions. The network is trained using feed-forward back propagation by repeatedly feeding training cases through the network. Based on the error in classification, the weights are updated backwards through the network. This process is repeated until the weights are sufficiently stable.

SVM: In the classification setting SVM attempts to find hyperplanes in the input space that best separates classes of the target variable. The hyperplane will be chosen such that the distance of the nearest points for the different classes to the hyperplane is a maximum.

5   Results and conclusion

Optimal tuning constants were found for all the techniques before they were compared with one another. Figure 2 shows the results for the 2 simulated data sets discussed is section 2. It can be seen that MARS performed well in both cases with support vector machines performing well on the second data set (right panel of figure 2). Note that SIMCA, which is currently the preferred method, did not perform as well. Figure 3 shows the results for the wine data set. MARS again performed well (based on the adapted accuracy) with SVM also giving good results. SIMCA performed worse than all the other techniques.
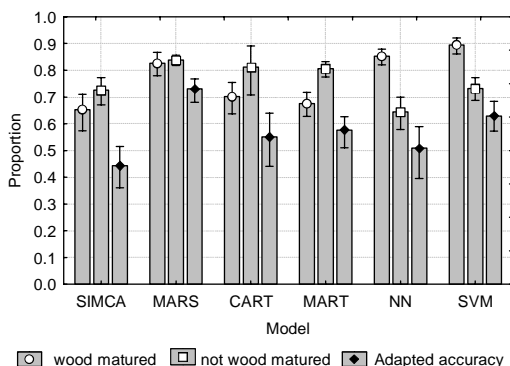
FIGURE 3. Results for the wine data set.

The techniques included in this study in general performed better than SIMCA (which is the current standard for NIR calibration). MARS overall gave the best results for all the data sets, with SVM also giving good results. Based on comments in the literature, much was expected of the boosting method MART, but it was generally outperformed by MARS and SVM.

References

Hastie, T., Tibshirani, R., Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference and Prediction.* Springer Series in Statistics.

Padin, P.M., Pena, R.M., Garcia, S., Iglesias, R., Barro, S., Herrero, C. (2001). Characterization of Galician (N.W. Spain) quality brand potatoes: a comparison study of several pattern recognition techniques. *Analyst*, **126**, 97-103.

Roggo, Y., Duponchel, L., Ruckebusch, C., Huvenne, J-P.](2003). Statistical tests for comparison of quantitative and qualitative models developed with near infrared spectral data. *Journal of Molecular Structure*, **654**, 253-262.

Wold, Svante (1976). Pattern recognition by means of disjoint principal component models. *Pattern recognition*, **8**, 127-139.

Identifying important input variables by applying alignment in kernel Fisher discriminant analysis

N. Louw and S.J. Steel[1]

[1] Department of Statistics and Actuarial Science, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa

Abstract: Kernel Fisher discriminant analysis (KFDA) is a recent nonlinear extension of discriminant analysis. We apply KFDA to a South African coronary heart disease risk factor data set. A new measure of variable importance in KFDA is introduced, and successfully used to rank the risk factors in order of importance.

Keywords: dimension reduction, kernel methods, variable importance.

1 Introduction

Since the introduction of support vector machines during the early 1990s, kernel based methods have become popular tools for classification and regression in the machine learning community. This trend is also evident in statistics, especially since kernel methods frequently outperform traditional statistical procedures (cf. Hastie et al., 2001). Examples of popular kernel methods are kernel principal component analysis, kernel logistic regression, and kernel Fisher discriminant analysis (KFDA). These methods are characterised by transformation of the input data to a high dimensional feature space, followed by application of the technique in question to the transformed data. Provided application of the technique requires only calculating inner products between pairs of input vectors, the so-called kernel trick obviates explicit calculations in the feature space. The focus in this paper is on KFDA, an extension of linear discriminant analysis. KFDA was introduced by Mika et al. (1999), and it has since been found to perform very well compared to traditional statistical classification procedures. Although the KFDA algorithm usually classifies quite accurately, it does not provide a natural way of determining the relative importance of the input variables. In this paper we therefore apply the concept of alignment to a practical two group classification problem to rank the input variables in terms of their ability to separate the two groups. This suggests a natural procedure for dimension reduction, and we see that for our problem the accuracy of KFDA classification is indeed slightly improved if the full set of input variables is replaced by a subset selected from the alignment values.

In Section 2 we provide a very brief overview of KFDA. Section 3 contains a discussion of the concept of alignment, and we argue that alignment can be used to define a measure of variable importance. We describe the data analysis and results in Section 4.

2   Kernel Fisher discriminant analysis

Consider the following generic two-group classification problem. We observe a binary response variable $Y \in \{-1, +1\}$, together with classification or input variables $X_1, X_2, \cdots, X_p$. These variables are observed for $N = N_1 + N_2$ sample cases, with the first N_1 cases coming from population 1 and the remaining N_2 cases from population 2. The resulting training data set is therefore $\{(\vec{x}_i, y_i), i = 1, 2, \cdots, N\}$. Here, $\vec{x}_i$ is a p-component vector representing the values of $X_1, X_2, \cdots, X_p$ for case i in the sample. Our purpose is to use the training data to determine a rule that can be used to assign a new case with observed values of the predictor variables in a vector $\vec{x}$ to one of the two classes. The KFDA classification rule is given by $sign\left\{b + \sum_{i=1}^{N} \alpha_i K(\vec{x}_i, \vec{x})\right\}$. Here, b and $\alpha_1, \alpha_2, \cdots, \alpha_N$ are quantities determined by applying the KFDA algorithm to the training data, while $K(\vec{x}_i, \vec{x})$ is a kernel function evaluated at $(\vec{x}_i, \vec{x})$. Two examples of popular kernel functions are the polynomial kernel, $K(\vec{x}_1, \vec{x}_2) = <\vec{x}_1, \vec{x}_2>^d$, where d is an integer, usually 2 or 3, and the Gaussian kernel, $K(\vec{x}_1, \vec{x}_2) = exp(-\gamma\|\vec{x}_1 - \vec{x}_2\|^2)$, where γ is a so-called kernel hyperparameter. We restrict attention to the Gaussian kernel in the remainder of the paper. For a more detailed discussion of KFDA, see for example Mika et al. (1999).

3   Alignment as a measure of variable importance

An important property of support vector machines is that the input vectors $\vec{x}_i$ appear in the algorithm only as arguments of the kernel function, i.e. we encounter these vectors only in the form $K(\vec{x}_i, \vec{x}_j), i, j = 1, 2, \cdots, N$. Evaluating $K(\vec{x}_i, \vec{x}_j)$ for $i, j = 1, 2, \cdots, N$, we are able to construct the so-called Gram matrix with ij-th entry $K(\vec{x}_i, \vec{x}_j)$. When a support vector machine is applied to a two-group classification problem, the Gram matrix contains all the information provided by the input vectors $\vec{x}_i$. Since $K(\vec{x}_i, \vec{x}_j)$ can be interpreted as a measure of the similarity between $\vec{x}_i$ and $\vec{x}_j$, Cristianini et al. (2002) argue that an ideal Gram matrix would be of the form $\vec{y}\vec{y}'$, where $\vec{y}$ is the N-component response vector with -1 in the first N_1 positions and +1 in the remaining N_2 positions. They define the concept of (empirical) alignment between a given Gram matrix $\mathbf{G} = [K(\vec{x}_i, \vec{x}_j)]$ and the ideal Gram matrix $\vec{y}\vec{y}'$ by

$$A(\mathbf{G}, \vec{y}\vec{y}') = \frac{< \mathbf{G}, \vec{y}\vec{y}' >_F}{\sqrt{< \mathbf{G}, \mathbf{G}' >_F < \vec{y}\vec{y}', \vec{y}\vec{y}' >_F}} \ , \tag{1}$$

where $< \mathbf{R}, \mathbf{S} >_F = trace(\mathbf{RS})$ is the Frobenius inner product between the symmetric matrices $\mathbf{R}$ and $\mathbf{S}$. These authors investigate the properties of the alignment, the most important for our purpose being that a large value of the alignment is desirable, since this will typically lead to the kernel method generalizing well, i.e. classifying new cases accurately.

Alignment can now be used to define a quantity that reflects the importance of an input variable in KFDA as follows. Consider the Gaussian kernel, and let $K_r(\vec{x_i}, \vec{x_j}) = exp[-\gamma(x_{ir} - x_{jr})^2]$ with corresponding Gram matrix $\mathbf{G}_r, r = 1, 2, \cdots, p$. These are the Gram matrices obtained by evaluating the kernel function on a single coordinate of the input vectors at a time. The importance of variable X_j can now be judged in terms of the alignment of $\mathbf{G}_j$ with the ideal Gram matrix $\vec{y}\vec{y}'$, i.e. by calculating $A(\mathbf{G}_j, \vec{y}\vec{y}')$. A large value of $A(\mathbf{G}_j, \vec{y}\vec{y}')$ would imply that X_j is an important input variable in the sense that it contributes significantly to separating the two populations under consideration.

Several points deserving further attention have to be made regarding this proposal to use $A(\mathbf{G}_j, \vec{y}\vec{y}')$ as a measure of individual variable importance. (i) The quantity $A(\mathbf{G}_j, \vec{y}\vec{y}')$ depends on the values of the kernel hyper-paramters. For the Gaussian kernel there is only a single hyperparameter, viz. γ. A decision has to be made regarding the value of γ to use when calculating $A(\mathbf{G}_j, \vec{y}\vec{y}')$. We found empirical evidence in simulation experiments in favour of using a fixed value of γ, for example $\gamma = 1$. (ii) What about other more well known measures than $A(\mathbf{G}_j, \vec{y}\vec{y}')$ to describe the importance of the input variables, for example correlation coefficients? In this regard it should be borne in mind that by using a kernel function one is able to exploit highly nonlinear relationships between the input variables and the binary response. It seems that a measure such as $A(\mathbf{G}_j, \vec{y}\vec{y}')$ is able to capture such nonlinear relationships, something which will be difficult if instead we calculate correlation coefficients. (iii) A further question that arises is whether the measure of variable importance can be used for effective dimensionality reduction. This would of course have the advantage that only a subset of the original input variables need to be used in further analyses and it may even lead to better classification performance of the resulting rule. The crucial issue in this regard is how to decide on the number of input variables to retain. This question is similar to the problem of deciding on the number of principal components or factors to use when performing a principal component or factor analysis. One strategy could be to use a scree plot of the ranked alignment values, and this possibility is explored in Section 4.

4   Analysis of the data set, and results

The data that were analyzed were collected as part of a study on risk factors in coronary heart disease that was conducted in South Africa. We

TABLE 1. Input variables ranked according to alignment values

X_2	X_5	X_3	X_9	X_8	X_7	X_1	X_4	X_6
0.154	0.113	0.107	0.062	0.056	0.056	0.046	0.044	0.042

consider $p = 9$ input variables and a binary response variable measured for each of 462 individuals. For the response variable, $y = +1$ indicates that the particular individual suffers from coronary heart disease, while $y = -1$ implies a control case. There were 160 diseased individuals and 302 control cases. The input variables, $X_1, X_2, \cdots, X_9$, were: systolic blood pressure, cumulative tobacco use, low density lipoprotein cholesterol, adiposity, family history of heart disease, an index of type-A behaviour, obesity, current alcohol consumption, and age.

We started our analysis by calculating $A(\mathbf{G}_j, \vec{y}\vec{y}')$, using $\gamma = 1$, for each of the input variables. This gave the values in Table 1, where we have ranked the variables according to alignment.

From Table 1 we see that the three most important input variables are cumulative tobacco use, family history of heart disease, and low density lipoprotein cholesterol. The decrease in alignment to the next variable, age, seems quite large, and we conjecture that the first three input variables may be sufficient to separate the two groups if a Gaussian kernel is used. Similar results were obtained for other constant values of γ. It is interesting to note that X_2, X_5, X_3 and X_9 are selected by a stepwise logistic regression procedure (see Hastie et al., 2001).

In an attempt to decide on the number of variables to retain, a scree plot of the ranked alignments was constructed (see Figure 1). It is clear that a levelling off in alignment occurs from X_9 onwards. This suggests using only the variables X_2, X_5 and X_3 in the KFDA rule.

To evaluate the classification performance of the KFDA rule based on different sets of variables, we repeated the following procedure 100 times. We randomly divided the 160 data cases pertaining to the diseased individuals into a training set of 96 cases and a test set of 64 cases. A similar division of the 302 control data cases into sets of respective sizes 181 and 121 was done. We then performed 9 KFD analyses: using only X_2, using X_2 and X_5, using X_2, X_5 and X_3 (the model suggested by the scree plot), up to an analysis based on all 9 input variables. In each case the KFDA algorithm was applied to the combined training data cases, and thereafter used to classify the test set cases. Table 2 summarizes the average test errors that were obtained in this way. The lowest test error was for KFDA based on the three input variables identified as most important by the proposed alignment measure, and suggested by the scree plot. This provides an indication that using alignment to identify important variables and to reduce dimensionality, may indeed have some merit.
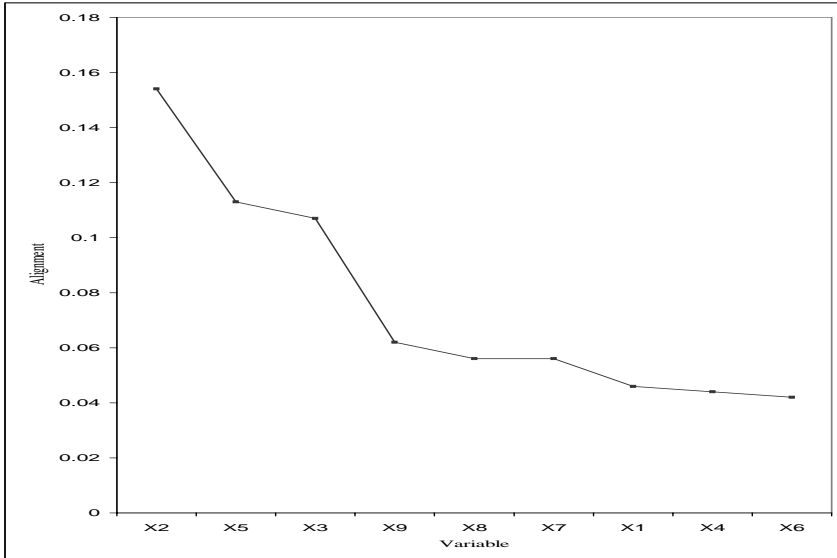
FIGURE 1. Scree plot of ranked alignment values

TABLE 2. Test errors for KFDA: successively adding more input variables

X_2	$+ X_5$	$+ X_3$	$+ X_9$	$+ X_8$	$+ X_7$	$+ X_1$	$+ X_4$	$+ X_6$
0.318	0.314	0.301	0.313	0.323	0.327	0.327	0.329	0.317

References

Cristianini, N., Kandola, J., Elisseeff, A. and Shawe-Taylor, J. (2002): On kernel-target alignment. In: T. Dietterich, S. Becker and D. Cohn (Eds.) *Neural Information Processing Systems*, **14**. Cambridge: MIT Press.

Hastie, T., Tibshirani, R. and Friedman, J. (2001): *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Heidelberg: Springer.

Mika, S., Rätsch, G., Weston, J., Schölkopf, B. and Müller, K.R. (1999). Fisher discriminant analysis with kernels. In Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, (Eds.) *Neural Networks for Signal Processing*, **IX**, 41-48. IEEE.

Statistical modelling for the time projection chamber signal processing: how can statistics improve detector performances?

Sara Maniero[1], Laura Ventura[1], Francesco Pietropaolo[2] and Sandro Ventura[2]

[1] Dept. of Statistics, via C. Battisti 241, Padova, Italy (`ventura@stat.unipd.it`)
[2] Dept. of Physics and INFN, via Marzolo 8, Padova, Italy (`sandrov@pd.infn.it`)

Abstract: The aim of this contribution is to investigate on how to improve the extraction of physical information from the signals coming from a liquid Argon (LAr) time projection chamber (TCP), a particle detector technique characterized by good tracking and energy measurement capabilities. We present here the results obtained from the analysis of test pulse data, i.e. the electronic impulses that, on purpose of calibration and testing, stimulate the electronics simulating a known charge value as if it was released by a particle within the LAr. Starting from the analysis of those calibration data, we focused on getting a better modelling of the electronic noise, which results far from a white noise process. As a subsequent step, we identified a more suitable theoretical analytical function to perform the nonlinear least-squares fit of the signal, used to recover the parameters which are relevant for the physical analysis.

Keywords: Autocorrelation, Integrated models, Least-squares fit, Nonlinear regression.

1   Introduction

The ICARUS project (Rubbia, 1977; ICARUS collaboration, 2001) is based on a large mass LAr TPC aimed to search for rare events, such as neutrino interactions or proton decay. The construction of the detector and the complete readout system of the LAr TPC are described for example in Amerio *et al.* (2003). Such a readout is based on the collection of the ionization electrons which are released when a charged particle travels through the LAr. The resulting signals on each channel are digitized and stored as waveforms which carry both spatial (time coordinate) and charge (area) information about the collected electrons. Reconstruction of a given particle event requires the measure of this charge on every different channel along the track path (depending on the particle energy and on the type of interaction the total number of channel outputs to be considered can go from few channels to many thousand) according to the specific channel characteristics (amplification factor, signal shaping time constants). Moreover the output signal
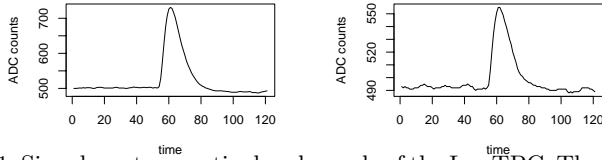
FIGURE 1. Signals on two particular channels of the Lar TPC. The peak indicates the charge released by a particle in the LAr.

shape is directly related to the ionization charge space distribution, which mainly depends on the track angle with respect to the readout plane. Such a variability of the signal, together with the non-negligible though unavoidable level of noise from the electronics, is a limiting factor on the choice of the analysis methods, as it seriously affects the possibility of using basic deconvolution techniques, and weakens the effectiveness of other considered statistical approaches, such as wavelet analysis (Polchlopek *et al.*, 2002) or neural network approach (ICARUS collaboration, 1995). For illustration purposes, Figure 1 presents two typical signals for two particular readout channels, showing both the electonic noise baseline, which depends to the specific channel characteristics, and the peak corresponding to a collected charge signal, which is the physical quantity of interest to be *modeled*.

Since one event is charaterized by all the signals resulting on each channel in one particular time interval (one event can involve even 2000 channels), to study the LAr TPC signal we focused on a simple procedure which can be applied automatically to each channel. So, on a first issue, the ICARUS analysis procedures simplified the extraction of the physical quantities from observed data, assuming a *nearly* white noise process for the electronic noise and performing a least-squares fit of the peak signal, i.e. the ionization charge, using a well-specified theoretical analytical function of the form $f(t; \beta)$, where t denotes time and β an unknown vector of parameters.

In this paper we present a short analysis of the essential features of a statistical approach to model LAr TCP signal. First of all, we adjust the model for the electronic noise using standard procedures of time series analysis. Accordingly, we propose a new theoretical analytical function to model the signal where the charge collection occurs, ponting out that $f(t; \beta)$ can limit the goodness of the fit of the peak. Although the proposed procedure may seems computationally intensive and time consuming, it is encompassed by the potential of modern statistical environments such as R.

2 Modelling the electronic noise

The data set used to study our models were recorded during the calibration test of the detector in the first technical run described in Amerio *et al.* (2003). Those are the first available data coming from the detector in its final working condition, so they allowed us to focus on the actual electronic noise behaviour. In particular, Figure 2 (a) gives an example of the electronic noise $\{\epsilon_t\}$ recorded by a specific channel. Given the tolerances in the electronics components and the differences in circuit and wiring layout, each readout channel actually shows a specific behavior and carries a possibly different noise figure. When designing the signal analysis procedure this has to be taken into account, avoiding as much as possible a direct dependency of the procedure on specific channel characteristics.

By looking to the correlogram (ACF) and to the partial ACF in Figure 2 (b) and (c), respectively, we note that these plots do certainly not agree with a white noise process. In particular, they indicate the presence of a trend, around which certain seasonal variations are apparent. The seasonal pattern is in this context very complex since it is the sum of several causes: mains power (very long period), feeders (short period), surrounding electric appliance interferences, LAr motion in the detector and mechanic vibrations. Also the classical Ljung-Box test statistic (see Wei, 1990, sec. 7.5) for examining the null hypothesis of independence in the time series indicates evidence against the null hypothesis. Inference based on $\{\epsilon_t\}$ generally makes ordinary least-squares estimation of β inefficient and standard errors of the estimates can be severely biased.

In practice, all the series $\{\epsilon_t\}$ observed in all the channels are non-stationary. In order to fit a stationary model, it is necessary to remove non-stationary sources of variation. Several methods of prefiltering the signals have been explored. For example, we investigated seasonal autoregressive integrated moving average models, but it turned out very difficult to adopt the same model on more of 1000 series involved in one single event. In view of this, we focused for a simpler method, which can be automatically implemented in all the channels.

One simple possibility is to difference the series and such a model is called an integrated model. In all the channels considered, it turned out that if we replace the electronic noise ϵ_t simply by $\nabla \epsilon_t = \epsilon_t - \epsilon_{t-1} = \epsilon_t^*$, then ϵ_t^* performs as a white noise process. The use of ϵ_t^* instead of ϵ_t has two consequences. From a theoretical point of view, the hypotheses necessary for a nonlinear least-squares fit of the event of interest are respected. From a practical point of view, handling the differenced series makes computational methods faster and reduces the number of signal samples around the peak to be stored, whilst insuring a good reconstruction of the charge signal.
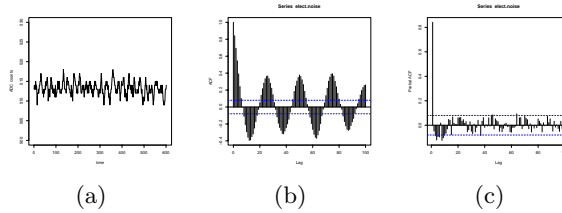
(a) (b) (c)

FIGURE 2. (a) A digitised electronic signal on a channel of the Lar TPC; (b) The ACF of the electronic signal; (c) The PACF of the electronic signal.

3 Fitting the charge signal

The aim of this section is to model the response of the detector readout to a collected charge. Indeed, the fit of the peak is the first step on the data analysis, since by integrating the fits we obtain a measure of the charge released by a particle in the LAr. We assume a model of the form

$$y_t = b(t, \beta) + \epsilon_t^* \ ,$$

where y_t denote the differentiated signal, $b(\cdot)$ denotes a deterministic component of the series, and ϵ_t^* is the error term. Function $b(t, \beta)$ is a nonlinear function of the time t and a vector of parameters β when the error is additive. Two different deterministic functions have been considered: the first one is simply given by $f(t; \beta) - f(t - 1; \beta)$, and the second one is a new proposal.

As in linear regression, parameter estimates are taken to be the values of β, which minimize the residual sum of squares $S(\beta) = \sum_{i=1}^{n} (y_t - b(t, \beta))^2$. Nonlinear regression requires calculation by iterative computer programs, which require initial estimates. Model goodness-of-fit may be examined using residuals (see Davison, 2003, chap. 10).

In our analysis, we found several improvements by using the new analytical function $b(t, \beta)$, with respect to the previous analysis procedure based on $f(t; \beta)$. Effectiveness of the new method has been verified through the evaluation of the electronics parameters (gain, linearity) which are needed to calibrate the charge response of every channel, showing the gained robustness of the fit against signal baseline fluctuations. A further qualification of the fitting procedure, which is underway, requires to apply the new charge estimation to a set of real particle tracks (such as illustrated in Figure 3), in order to explore the fit goodness over a full sample of the different possible signal shapes.

4 Final remarks

In this contribution we discuss how to improve the extraction of physical information from the signals coming from a liquid LAr TCP. In particular,
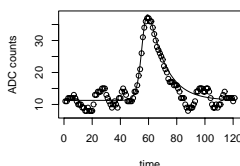
FIGURE 3. Fit based on $b(t, \beta)$ applied to a minimum ionizing particle signal.

we discuss a simple statistical proposal which is based on the use of differenced data, which presents the advantage to be applied automatically to each channel involved in one event. As a different charge estimation method we also considered a neural network based algorithm, but this approach didn't succeed mainly since it is not possible to build a statistical estimator whose value is intended as a meaningful guess for the unknown value of a parameter, or define a confidence level as in the normal best-fit procedures, and because it shows a strong dependency on the quality of the training example set, which in our case is quite difficult to qualify given the huge variability of the detector signal behavior. Another explored technique tried to exploit wavelet transforms to remove the noise from the signal. But although this technique resulted quite effective in reaching high data compression ratio, the charge estimation didn't perform better than the fitting procedures producing broader area distributions.

References

Amerio, S. *et al.* (2003). Design, construction and tests of the ICARUS T600 detector. *Submitted to Nucl. Instr. and Meth. in Phys. Res.* A.

Davison, A.C. (2003). *Statistical Models.* Cambridge University Press.

ICARUS Collaboration (1995). A neural network approach for the TPC signal processing. *Nucl. Instr. and Meth.* A, 356, 507.

ICARUS Collaboration (2001). The Icarus Experiment: A second-generation Proton decay experiment and Neutrino Observatory at Gran Sasso Osservatory - Cloning of T600 Modules to reach the design sensitive Mass. *LNGS-EXP*, 13/89, add. 2/01.

Polchlopek, W. *et al.* (2002). Wavelet transform compression and denoising in real-time system (Proposal for the ICARUS DAQ System). *Technical Memo*, ICARUS-TM/02-12.

Rubbia, C. (1977) The Liquid-Argon time projection chamber: A new concept for Neutrino Detector. *CERN-EP*, 77-08, Geneve (CH).

Wei, W.S. (1990). *Time Series Analysis*, Addison-Wesley.

Assessing the effect of a teaching program on breast self-examination in a randomized trial with noncompliance and missing data

Alessandra Mattei[1] and Fabrizia Mealli [1]

[1] Dipartimento di Statistica "G. Parenti" - Università di Firenze
   Viale Morgagni 59, 50134 Firenze - Italy
   mattei@ds.unifi.it   mealli@ds.unifi.it

Abstract: Many randomized studies suffer from noncompliance and missing data. We present an extended framework for the analysis of data from such experiments. We use an instrumental variables approach to link intention-to-treat effects to treatment effects and we adopt a Bayesian approach for inference and sensitivity analysis. This framework is illustrated in the context of a randomized trial of breast self-examination.

Keywords: Bayesian analysis; Rubin Causal Model; Instrumental variable; Non-compliance; Non-ignorable missing data; Intention-to-treat; Sensitivity analysis.

1   Introduction

In this paper we investigate the effect of the receipt of a treatment in the context of a randomized trial which suffers from noncompliance and missing outcomes. Specifically, we consider the consequences of two exclusion restrictions on the effect of assignment: an econometric exclusion restriction that disallows, for a specific subpopulation, direct links between assignment and outcome other than through the effect of assignment on the treatment received, and a response exclusion restriction, which requires that subjects who always comply with their assignment (whether it is to the new or control treatment) are not affected in their response behavior by their assignment (Mealli et al., 2004). Our Bayesian approach allows for the comparison of results based on different combinations of these assumptions, thereby assessing sensitivity to their violations.
We apply these methods to a randomized trial on Breast-Self-Examination (BSE), which was affected by the two sources of bias mentioned above.

2   The randomized trial on breast self-examination

In this paper, we consider a randomized trial of Brest self-examination. In this study, two BSE teaching methods were compared, a 'standard' treatment of receiving mailed information only, and an 'enhanced' treatment of

additional attendance in a self-exam course. The study was conducted over a 3-year period (1988-1990) at the Oncologic Center of the Faenza Health District in Italy (see previous analysis by Ferro et al., 1996 and Mealli et al., 2004). In order to address the noncompliance and missing data problems let us introduce some notation. For each individual i $(i = 1, \ldots, N)$ who partecipates in the study, let Z_i^{obs} represent their treatment assignment with $Z_i^{\text{obs}} = 1$ for new and 0 for standard treatment. In addition, let $D_i(z)$ be an indicator for the treatment received, given assignment z, and let $D_i^{\text{obs}} = D(Z_i^{\text{obs}})$ be the actual treatment received, where $D_i^{\text{obs}}(0) = 0$, as women assigned to the standard treatment had non access to the training course. Similarly, define $Y_i(z)$ as the potential outcome, given assignment to treatment level z, and let $Y_i^{\text{obs}} = Y(Z_i^{\text{obs}})$ be the actual outcome observed. Lastly, let $R_i(z)$ represent the potential response indicator (1 if a subject responds to the post-test questionnaire, 0 for non-responders), given treatment z, and let $R_i^{\text{obs}} = R(Z_i^{\text{obs}})$ represent the actual response indicator. In addition, a vector of pre-treatment variable, $\mathbf{X}_i^{\text{obs}}$ is observed per subject. In our application, we consider only two covariates: X_{i1}^{obs}, a binary indicator for previous BSE practice and X_{i2}^{obs}, a binary indicator of good knowledge of breast pathophysiology.

The randomization of assignment guarantees that the pretreatment variables being closely balanced in the two subsample defined by assignment. The randomization does not, however, imply that the pretreatment variable are balanced in the subsamples defined by the actual treatment status. This imbalance suggests that we cannot simply compare outcomes by treatment status to obtain credible estimates of the effect of the new teaching program.

3   Modeling compliance and response behavior

In this section we focus on defining the causal effect of interest, the effect of the new, enhanced training class on BSE practice. Throughout this analysis we will make the Stable Unit Treatment Value Assumption (SUTVA) that there is interference between neither units nor different versions of the treatment.

Let U_i represent the treatment woman i would receive if assigned to the active treatment $(U_i = D_i(1))$. If $U_i = 1$, the woman i is a 'complier'; in contrast, if $U_i = 0$, the subject i is a 'never-taker'. For this experimental setting, this compliance status U_i can be viewed as a covariate which is observed only for women with $Z_i^{\text{obs}} = 1$; by randomization, however, it is guaranteed to have the same distribution in both treatment arm. Let $\mathbf{U}$ and N_u be, respectively, the N component vector with ith element U_i and the number of units of type u, $u = 0, 1$. In addition, let $\mathbf{Y}$ be the $N \times 2$ matrix of potential outcomes with ith row equal to $(Y_i(0), Y_i(1))$. Using this notation, the ITT $= \sum_{i=1}^{N} [Y_i(1) - Y_i(0)]/N$ effect of assignment on the

outcome can be defined as the weighted average

$$\text{ITT} = \frac{N_1}{N}\text{ITT}_1 + \frac{N_0}{N}\text{ITT}_0 \tag{1}$$

where, for $u \in \{0, 1\}$, $\text{ITT}_u = \sum_{i:U_i=u}[Y_i(1)-Y_i(0)]/N_u$ is the average ITT effect of Z on Y for each of the two sub-populations defined by compliance behavior, and N_u/N is the weight assigned to ITT_u.

Random assignment of the treatment implies that $\Pr(Z_i|D_i(0), D_i(1), Y_i(0), Y_i(1), \mathbf{X}_i^{\text{obs}}) = \Pr(Z_i)$. As conditioning on pretreatment variables assignment remains ignorable (Rubin, 1978), in general, we only require:

ASSUMPTION 1 *(Ignorability of treatment assignment)*

$$\Pr(Z_i|D_i(0), D_i(1), Y_i(0), Y_i(1), \boldsymbol{X}_i^{\text{obs}}) = \Pr(Z_i|\boldsymbol{X}_i^{\text{obs}}). \tag{2}$$

Concerning the response behavior, we assume that potential outcomes are independent of the missing indicator given observed covariates conditional on the compliance status and the assignment levels, that is:

ASSUMPTION 2 *(Latent Ignorability)*

$$R_i \perp Y_i | Z_i, \boldsymbol{X}_i^{\text{obs}}, U_i. \tag{3}$$

We also consider, but do not necessarily impose, two additional assumptions; two exclusion restrictions on the effect of assignment.

ASSUMPTION 3 *(Outcome exclusion restriction for never-takers)*

$$Y_i(Z_i) \perp Z_i | \boldsymbol{X}_i^{\text{obs}}, U_i = 0. \tag{4}$$

This assumption implies that $\Pr(Y_i(1)|\mathbf{X}_i^{\text{obs}}, U_i = 0) = \Pr(Y_i(0)|\mathbf{X}_i^{\text{obs}}, U_i = 0)$, so that within subpopulation of never-takers with the same values of cavariates, the distributions of the two potential outcomes $Y_i(0)$ and $Y_i(1)$ are the same.

When the outcomes are not observed for all units, since the compliance status is partially missing, latent ignorability is not sufficient to identify the ITT effect for compliers. To address this complication, Mealli et al. (2004) propose the following assumption:

ASSUMPTION 4 *(Response exclusion restriction for compliers)*

$$R_i(Z_i) \perp Z_i | \boldsymbol{X}_i^{\text{obs}}, U_i = 1. \tag{5}$$

This assumption implies that compliers have the same response behavior irrespective of the treatment arm they are assigned to.

We regard the two assumption 4 and 3 as possibly controversial, and we will investigate their consequences in some detail.

In order to relax fully one or both exclusion restrictions, we impose a parametric form of the likelihood function and using a relatively diffuse but proper prior distribution.

4   Parametric models

We model the conditional distribution of the compliance status U given the pretreatment variables $\mathbf{X}$ and the conditional distributions of the potential response indicator R and the potential outcome Y, given $\mathbf{X}$ and U. As all the variables of interest are dichotomous, we assume that their distribution have a logistic regression form:

$$\pi_i^U = \Pr\big(U_i = 1\big|\mathbf{X}_i = \mathbf{x}; \alpha\big) = \frac{\exp(\alpha_0 + \alpha_1'\mathbf{x})}{1 + \exp(\alpha_0 + \alpha_1'\mathbf{x})} \tag{6}$$

$$\pi_{iuz}^R = \Pr\big(R_i = 1\big|U_i = u, Z_i = z, \mathbf{X}_i = \mathbf{x}; \beta_{uz}\big) = \frac{\exp(\beta_{uz0} + \beta_{uz1}'\mathbf{x})}{1 + \exp(\beta_{uz0} + \beta_{uz1}'\mathbf{x})} \tag{7}$$

$$f_{izu}(1) = \Pr\big(Y_i = 1\big|U_i = u, Z_i = z, \mathbf{X}_i = \mathbf{x}; \gamma_{uz}\big) = \frac{\exp(\gamma_{uz0} + \gamma_{uz1}'\mathbf{x})}{1 + \exp(\gamma_{uz0} + \gamma_{uz1}'\mathbf{x})} \tag{8}$$

The full parameter vector, denoted by θ, has 27 elements. In the application in this paper, we impose prior equality of some slope coefficients: $\beta_{u11} = \beta_{u01}$, $\beta_{u12} = \beta_{12}$, $\gamma_{u11} = \gamma_{u01}$ $\gamma_{u12} = \gamma_{u02}$, for $u = 0, 1$, reducing the number of parameters to 19.

For inference, we consider the Markov chain algorithm, a variant of the Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings, 1970), which uses the Data Augmentation method of Tanner and Wong (1987).

As in Hirano et al. (2000), we use a relatively diffuse proper prior distribution with a simple conjugate form:

$$p(\theta) \propto \prod_{i=1}^{N} \times \prod_{u,z,r} \Big(\big(\pi_i^U\big)^u \big(1 - \pi_i^U\big)^{(1-u)} \big(\pi_{iuz}^R f_{iuz}(Y_i)\big)^r \big(1 - \pi_{iuz}^R\big)^{(1-r)}\Big)^{2.5/N} \tag{9}$$

5   Results and conclusions

In Table 1, summary statistics of the posterior distribution of the estimands of interest are presented under the four combinations of the two exclusion restrictions.

We find plausible to impose the response the exclusion restriction for compliers and relax the outcome exclusion restriction for never-takers. Therefore, we focus on the third block of columns in Table 1. The marginal distributions of the subpopulation ITT effects suggest that the effects for compliers and never-takers are very different. Examining their joint distribution in Figure 1, we see that the effects are somewhat negatively correlated. Specifically, we find a quite strong negative ITT effect for never-takers and

TABLE 1. Summary statistics: posterior distributions

Resp. Excl. Res. Compliers	Yes		NO		Yes		NO	
Excl. Res. Never-takers	Yes		Yes		NO		NO	
Estimand	Mean	sd	Mean	sd	Mean	sd	Mean	sd
ITT_c	-0.040	(0.050)	-0.008	(0.054)	0.058	(0.117)	0.075	(0.118)
ITT_n	0	0	0	0	-0.179	(0.228)	-0.141	(0.245)
ITT	-0.022	(0.028)	-0.004	(0.030)	-0.047	(0.048)	-0.020	(0.067)
$\Pr(R_i(1)=1|U_i=1)$	0.796	(0.030)	0.790	(0.031)	0.793	(0.031)	0.789	(0.031)
$\Pr(R_i(0)=1|U_i=1)$	0.796	(0.030)	0.890	(0.100)	0.793	(0.031)	0.814	(0.170)
$\Pr(R_i(1)=1|U_i=0)$	0.419	(0.041)	0.418	(0.042)	0.416	(0.042)	0.417	(0.041)
$\Pr(R_i(0)=1|U_i=0)$	0.541	(0.072)	0.431	(0.138)	0.543	(0.074)	0.518	(0.219)

a small and not much significant positive ITT effect on BSE practice for compliers. Concerning the response behavior, this model gives a plausible figures for the response probabilities: per assigned treatment level, never-takers have lower response rates than compliers. In addition, never-takers have a lower response rate if assigned to the new treatment arm than if assigned to the standard treatment.

Our analysis does not provide evidence that the overall ITT effect arises entirely or even largely from the effect of the training course on BSE techniques.
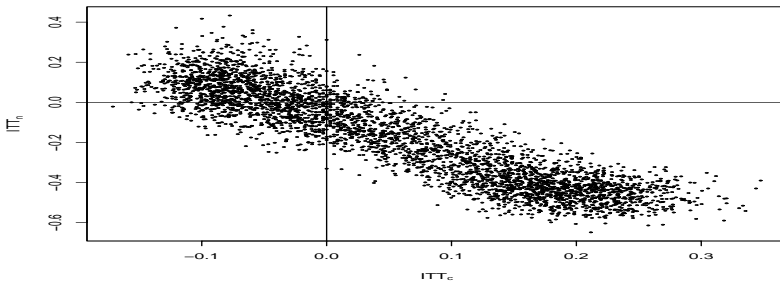
FIGURE 1. *Simulation scatterplot of the joint posterior distribution of ITT_c and ITT_n in the model with only response exclusion restriction for compliers.*

References

Hirano, K., Imbens, G.W., Rubin, D.B. and Zhou, X.H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, **1**, 69-88.

Mealli, F., Imbens, G.W., Ferro, S. and Biggeri, A. (2004). Analyzing randomized trial on breast self-examination with noncompliance and missing outcomes. *Biostatistics*, **5**, 207-222.

Variance free model for two-way layout with interaction

Joao T. Mexia[1], Stanisław Mejza[2]

[1] Departmento de Matematica, Universidade Nova de Lisboa, Quinta da Torre, 2825 Monte da Caparica, Portugal
[2] Department of Mathematical and Statistical Methods, Agricultural University, Wojska Polskiego 28, PL-60-637 Poznań, Poland

Abstract: The paper deals with an application of variance free model approach to the analysis of the two-factor experiments carried out in completely randomized design. Moreover, an estimation and testing hypotheses concerning main and interaction effects are considered. The application of the experiment considered in the genetics is discussed as well.

1 Introduction

Let us consider a two-factor experiment in which the first factor A occurs at s levels (treatments) $A_1, A_2, ..., A_s$, while the second factor B occurs at t levels (treatments) $B_1, B_2, ..., B_t$. Moreover, let us assume that the experimental material is homogeneous. Then the completely randomized design is appropriate for that structure. It means that all st treatment combinations $(A_i B_j)$ we can randomly arrange on the experimental units. The usual inference from this kind of experiments is well known and descried in many monographs.

Let us assume that the k-th replication of the observation y_{ijk} concerning the (i,j)-th treatment combinations $(A_i B_j)$ is modelled as follows:

$$y_{ijk} = \gamma_{ij} + e_{ijk}, \quad i = 1, 2, ..., s, \ j = 1, 2, ..., t, \ k = 1, 2, ..., n, \quad (1)$$

where γ_{ij} denotes the expected value of the trait observed on the (i,j)-th treatment combinations $(A_i B_j)$, n denotes the number of the $(A_i B_j)$ replications and finally, e_{ijk} denotes the error. It will be assumed that $e_{ijk} \sim N(0, \sigma^2)$ for all i, j, k. The efects of the (i,j)-th treatment combinations $(A_i B_j)$ can be expressed as:

$$\gamma_{ij} = \gamma_{..} + (\gamma_{i\cdot} - \gamma_{..}) + (\gamma_{\cdot j} - \gamma_{..}) + (\gamma_{ij} - \gamma_{i\cdot} - \gamma_{\cdot j} + \gamma_{..}), \quad (2)$$

where $\gamma_{..}$ denotes the general mean, $\alpha_i = (\gamma_{i.} - \gamma_{..})$ - the effect of the i-th level effect of factor A, $\beta_j = (\gamma_{.j} - \gamma_{..})$ - the j-th level effect of factor B, $\omega_{ij} = \gamma_{ij} - \gamma_{i.} - \gamma_{.j} + \gamma_{..}$ the interaction effect of the i-th level effect of factor A with the j-th level effect of factor B. We use the classical dot notation for means.

Now let us define the general hypotheses that to be veryfied by the two-factor experiment. The hypotheses can be expressed as:

$H_{0\alpha}$: $\alpha_i = 0$, for all i, $i=1,2,...,s$,
$H_{0\beta}$: $\beta_j = 0$, for all j, $j=1,2,..., t$,
$H_{0\alpha\beta}$: $\omega_{ij} = 0$, for all i,j, $i=1,2,...,s, j=1,2,..., t$.

The above hypotheses can be veryfied by using standard analysis of variance technique.

2   Variance free model

Let us assume that on each experimental unit we observe two continuous traits (random variables) say (X, Y) and let their joint distribution be normal. Moreover, let us take n observations on each treatment combination (A_iB_j), $(x_{ij1}, y_{ij1}), ..., (x_{ijn}, y_{ijn})$. The inference concerning treatment (factor) effects can be based on these traits independently. But this is correct only when the traits are uncorrelated (independently distributed under normality). However, many times the traits are correlated and then it is necessary to take this fact into account in inference from the experiment considered. Hence, in this paper we propose a way to infer on treatment effects taking into account possible correlation between traits. The analysis proposed is based on the correlation coefficients. Another approach could be based, for instance, on MANOVA techniques.

Let ρ_{ij}, $i=1,2,..., s$, $j=1,2,...,t$ be the correlation coefficient for the (i,j) treatment combination (A_iB_j) and let r_{ij} be its estimator. Then using the transformation (cf. Kendal and Stuart, 1958, Mexia, 1990)

$$z_{ij} = 0.5\sqrt{n-3}\ ln((1+r_{ij})/(1-r_{ij})) \qquad (3)$$

we obtain $z_{ij} \sim N(\mu_{ij}, 1)$ where $\mu_{ij} = 0.5\sqrt{n-3}\ ln((1+\rho_{ij})/(1-\rho_{ij})) + (\rho_{ij}\sqrt{n-3})/(2(n-1))$, $i=1,2,...,s, j=1,2,...,t$.

We use this transformation when the number of treatment combination is quite large. Then $(\rho_{ij}\sqrt{n-3})/(2(n-1))$ is proportionally small with respect to the first part of μ_{ij}. Hence, in further considerations we will

assume that $z_{ij} \sim N(\tilde{\mu}_{ij}, 1)$, with $\tilde{\mu}_{ij} = 0.5\sqrt{n-3}\, ln((1+\rho_{ij})/(1-\rho_{ij})) = c\, ln\phi_{ij}$, where $c = 0.5\sqrt{n-3}$, $\phi_{ij} = (1+\rho_{ij})/(1-\rho_{ij})$.,
Finally, expressing $\tilde{\mu}_{ij}$ in the same way as γ_{ij} in (2) we obtain the model

$$\tilde{\mu}_{ij} = \tilde{\mu} + \tilde{\alpha}_i + \tilde{\beta}_j + \tilde{\omega}_{ij}, \tag{4}$$

where $\tilde{\mu}$ is the general mean, $\tilde{\alpha}_i$, $\tilde{\beta}_j$ are the effects of factor A and B levels, while the $\tilde{\omega}_{ij}$ are the interaction effects.
Then we can express z_{ij} as

$$z_{ij} = \tilde{\mu} + \tilde{\alpha}_i + \tilde{\beta}_j + \tilde{\omega}_{ij} + \tilde{e}_{ij}, \tag{5}$$

where $\tilde{e}_{ij} \sim N(0,1)$.
Model (5) is called variance free model for two factor experiment carried out in completely randomized design.

To find the estimators of the treatment effect contrasts and interaction effect contrasts in model (5) we can use analysis of variance technique for two-factor experiment without replications. Let us note that all three hypotheses mentioned earlier are testable (variance is known).
A problem worth noticing is connected with the meaning of the hypotheses considered in the model (2) in relation to variance free model (5).
The hypothesis: $H_{0\alpha} : \tilde{\alpha}_i = 0$, for all i, is equivalent to

$H_{0\alpha} : \Pi_{j'=1}^{t} \phi_{ij'} = c_\alpha$, for all i, where $c_\alpha = (\Pi_{l=1}^{s}\Pi_{v=1}^{t}\phi_{lv})^{1/s}$.

Similarly, $H_{0\beta} : \tilde{\beta}_j = 0$, for all j is equivalent to

$H_{0\beta} : \Pi_{i'=1}^{s} \phi_{i'j} = c_\beta$ for all j, $c_\beta = (\Pi_{l=1}^{s}\Pi_{v=1}^{t}\phi_{lv})^{1/t}$.

Finally, let us consider the $H_{0\alpha\beta} : \tilde{\omega}_{ij} = 0$, for all i and j.
This hypothesis is equivalent to $H_{0\alpha\beta} : \phi_{ij} = c_{i,j}$, for all i and j

$$c_{i,j} = (\Pi_{v=1}^{t}\phi_{iv})^{1/t}(\Pi_{l=1}^{s}\phi_{lj})^{1/s}/(\Pi_{l=1}^{s}\Pi_{v=1}^{t}\phi_{lv})^{1/st} \tag{6}$$

Let us note that hypothesis (6) is an multiplicative version of the very well known Fisher condition for two classifications to be orthogonal.

3   Discussion

This kind of experiments is often performed in agricultural and biological research. Especially it is useful when we observe two correlated traits and one of them is easy to observe (measure) while to observe (measure) the second trait it is necessary to cut the plant or kill the animal. Hence, it is

recommended to identify two correlated traits, especially at the beginning of a research. Then by using parallel variance free aproach and usual analysis of variance aproach for two traits independently, we can compare the inference in both cases. Finally, the variance free aproach can be used to adjust the further inference based only on the trait that is easy to observe. The variance free model for two factor experiments was adapted to genetical experiments connected with breeding program. This kind of experiment is commonly performed by geneticists who are interested in selecting lines and strains of plants or animals for further breeding. The structure of the model used is similar to that of the two-way layout with interaction as considered here. In the first kind of such experiment, called *line x tester*, two sets of inbred lines are chosen and crosses among these lines are made. The first set of lines includes s chosen inbred lines, usually of unknown genetical value in the breeding program. The second set of lines includes t known and valuable lines called testers. Then, the line x tester system, involves crossing the s lines in the first group with each of the t testers.

The variance free approach to line x tester experiments is given in Mejza and Mexia (2002a).

In the second kind of such experiment, called *diallel cross experiment*, a set of s inbred lines is chosen and all possible crosses among these lines are made ($s=t$). It means that we can get sxs treatment combinations (crosses). The selecting process is based on the inference concerning main effects (called general combinig ability), interaction effects (called specific combining ability) and on additional effects called reciprocal effect.

The analysis of diallel cross experiment by variance free model approach is given by Mexia and Mejza (2002b).

References

Kendall, M., and Stuart, A. (1958). *The Advanced Theory of Statistics - Vol. I*, Charles Griffin

Mexia, J. T. (1990). Variance free models, *Trabalhos de Investigacao, Dept. of Mathematics, F.C.T., U.N.L.*, **2**.

Mejza, S., and J. T. Mexia . (2002a). Variance free model of line x tester experiments. *Statistical Modelling in Society. Proc. of the 17th International Workshop on Statistical Modelling Chania, Crete.* M. Stasinopoulos and G. Touloumi eds. 70-74.

Mexia, J. T., and Mejza, S. (2002b). Variance free model of diallel cross experiments. *15th Summer School in Biometry, CISTA, Brno, Czech Republik,* Hartmann and Pospisil, eds. 243-250.

Logit Model for TB in Europe (1995-2000)

Sandra Nunes[1], João Tiago Mexia[2] and Christoph Minder[3]

[1] Departamento de Matemática, Escola Superior de Tecnologia de Setúbal, Estefanilha, 2914-508 Setúbal, Portugal
[2] Departamento de Matemática, Universidade Nova de Lisboa, Quinta da Torre, 2829-516 Caparica, Portugal
[3] Department of Social and Preventive Medicine, University of Berne, Finken-hubelweg 11, 3012 Berne, Switzerland

Abstract: Structured Least Squares are used to adjust a two factor model for the logit of TB incidence. The factors considered were country and year. The data showed a regular decrease with time and a stable division of Europe in Eastern Europe, the Balkans and Western Europe.

1   Introduction

Our main goal is to obtain quantitative estimates for Tuberculosis (TB) in Europe.

To achieve this we applied logit model to the data for TB incidence. This data was organized per countries and covered the time span from 1995 to 2000. This data is available in Surveillance of Tuberculosis in Europe - Euro TB.

The Algorithm presented here allow us to obtain the estimates to our parameters α and β, when we just know the incidence of a disease for pairs (i, j).

2   Model and Algorithm

Let us assume that

$$y_{i,j} = logit p_{i,j} = \ln \frac{p_{i,j}}{1 - p_{i,j}} = \alpha + \beta \left(f_i + g_j \right) \tag{1}$$

with $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

Being $p_{i,j}$ the probability of an individual be infected with TB. And where f_i , $i = 1, \ldots, m$ and g_j , $j = 1, \ldots, n$ are two any unknown factors.

In our specific case we considered:

- f =exposure=country in study;

- g =susceptibility=year in study.

Let us put $x_{i,j} = f_i + g_j$; $i = 1, \ldots, m$; $j = 1, \ldots, n$ assuming as initial values for $x_{i,j}$ the following ones

$$x_{i,j}\left(\imath\right) = y_{i,\bullet} + y_{\bullet,j} - y_{\bullet,\bullet} \quad i = 1, \ldots, m \; ; j = 1, \ldots, n \tag{2}$$

where $y_{i,\bullet} = \frac{1}{n}\sum_{j=1}^{n} y_{i,j}$, $y_{\bullet,j} = \frac{1}{m}\sum_{i=1}^{m} y_{i,j}$ and $y_{\bullet,\bullet} = \frac{1}{m*n}\sum_{i=1}^{m}\sum_{j=1}^{n} y_{i,j}$.

Being $v_{i,j} = Var\left(y_{i,j}\right) \approx \frac{1}{N_{i,j} \times p_{i,j}}$; $i = 1, \ldots, m$; $j = 1, \ldots, n$, where $N_{i,j}$ represents the population in country i and in year j. Not to overload the notation let us put $q_{i,j} = \frac{1}{v_{i,j}}$; $i = 1, \ldots, m$; $j = 1, \ldots, n$.

So we may write that

$$S\left(\imath\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} q_{i,j}\left(y_{i,j} - \alpha - \beta\, x_{i,j}\left(\imath\right)\right)^2 = \tag{3}$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n} q_{i,j}\left(y_{i,j} - \alpha - \beta\left(f_i\left(\imath\right) + g_j\left(\imath\right)\right)\right)^2. \tag{4}$$

To lighten the notation let us put $\quad S\left(\imath\right) = S \quad and \quad x_{i,j}\left(\imath\right) = x_{i,j}$.

2.1   Zigzag Algorithm

We now describe the several steps of the algorithm applied.

Step 1 In the first step we minimize S in order to the parameters (α, β), using the initials values of $x_{i,j}$. From this minimization we obtained the following estimates:

$$\check{\alpha}\left(\imath\right) = \check{\alpha} = y_{\circ} - \check{\beta}x_{\circ} \quad and \quad \check{\beta}\left(\imath\right) = \check{\beta} = \frac{s_{x,y}}{s_{x,x}} \tag{5}$$

where

$$y_{\circ} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} q_{i,j}\, y_{i,j}}{q^+} \quad ; \quad x_{\circ} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} q_{i,j}\, x_{i,j}}{q^+} \tag{6}$$

with $q^+ = \sum_{i=1}^{m}\sum_{j=1}^{n} q_{i,j}$

and

$$s_{x,x} = \sum_{i=1}^{m}\sum_{j=1}^{n} q_{i,j}\left(x_{i,j} - x_{\circ}\right)^2 \tag{7}$$

$$s_{x,y} = \sum_{i=1}^{m} \sum_{j=1}^{n} q_{i,j} \left(x_{i,j} - x_\circ \right) \left(y_{i,j} - y_\circ \right). \tag{8}$$

Step 2 In this step we minimize

$$S = \sum_{i=1}^{m} \sum_{j=1}^{n} q_{i,j} \left(y_{i,j} - \check{\alpha} - \check{\beta} \left(f_i + g_j \right) \right)^2 \tag{9}$$

in order to the vectors $\underline{f}^m$ and $\underline{g}^n$. We will obtain the following system:

$$\left[\begin{array}{c|c} D_1 & Q \\ \hline Q^t & D_2 \end{array} \right] \left[\begin{array}{c} \underline{f}^m \\ \underline{g}^n \end{array} \right] = \underline{V}^{m+n} \tag{10}$$

where $Q = [q_{i,j}]$;

$$D_1 = D \left(\sum_{j=1}^{n} q_{1,j} , \ldots, \sum_{j=1}^{n} q_{m,j} \right) \tag{11}$$

$$D_2 = D \left(\sum_{i=1}^{m} q_{i,1} , \ldots, \sum_{i=1}^{m} q_{i,n} \right) \tag{12}$$

and the components of $\underline{V}^{m+n}$ are:

- $V_i = \frac{1}{\check{\beta}} \sum_{j=1}^{n} q_{i,j} \left(y_{i,j} - \check{\alpha} \right)$; $i = 1, \ldots, m$

- $V_{m+j} = \frac{1}{\check{\beta}} \sum_{i=1}^{m} q_{i,j} \left(y_{i,j} - \check{\alpha} \right)$; $j = 1, \ldots, n$.

Solving this system we will obtain the new values of f and g : $\check{f}_i \left(\imath \right)$, $i = 1, \ldots, m$, and $\check{g}_j \left(\imath \right)$, $j = 1, \ldots, n$, and consequently $\check{x}_{i,j} \left(\imath \right) = \check{f}_i \left(\imath \right) + \check{g}_j \left(\imath \right)$.

Step 3 In the third step we calculate

$$\tilde{S} \left(\imath \right) = \tilde{S} = \sum_{i=1}^{m} \sum_{j=1}^{n} q_{i,j} \left(y_{i,j} - \check{\alpha} \left(\imath \right) - \check{\beta} \left(\imath \right) \left(\check{f}_i \left(\imath \right) + \check{g}_j \left(\imath \right) \right) \right)^2 \tag{13}$$

where $\check{\alpha} \left(\imath \right)$, $\check{\beta} \left(\imath \right)$, $\check{f}_i \left(\imath \right)$, $i = 1, \ldots, m$, and $\check{g}_j \left(\imath \right)$, $j = 1, \ldots, n$, are the adjusted values obtained in cycle $\imath$.

Step 4 In this last step we carry out the standardization in order to keep unchanged the minimum and the maximum of $x_{i,j}$.

The values obtained from this standardization will be used in the next cycle if the value of $\tilde{S} \left(\imath \right)$ will not have stabilized.

3   Results and Conclusions

We analyzed the incidence of TB in fifty one European countries ($m = 51$) covering six years ($n = 6$), from 1995 to 2000.
After applying our algorithm we obtained the following results:

- The estimates for α and β :

$$
\begin{cases}
\check{\alpha} = -0.607712 \\
\check{\beta} = 0.918492
\end{cases}
. \tag{14}
$$

- In Figure 1 we present the values for the factors matching exposure f.

Country		Country		Country	
Georgia	2,61	Bulgaria	1,26	France	-0,10
Kazakhstan	2,45	Hungary	1,21	Finland	-0,13
Romania	2,32	Tajikistan	1,16	Ireland	-0,13
Kyrgyzstan	2,31	Yugoslavia	1,15	United Kingdom	-0,24
Russia	1,99	Turkey	1,09	Switzerland	-0,25
Latvia	1,93	Poland	1,08	Denmark	-0,34
Lithuania	1,89	Macedonia	1,03	Netherlands	-0,34
Bosnia-Herzeg.	1,85	Armenia	0,94	Luxembourg	-0,39
Turkmenistan	1,85	Slovakia	0,69	Israel	-0,45
Moldova	1,76	Slovenia	0,65	Greece	-0,45
Belarus	1,60	Spain	0,55	Italy	-0,47
Azerbaijan	1,59	Albania	0,54	Monaco	-0,72
Ukraine	1,59	Andorra	0,40	San Marino	-0,89
Uzbekistan	1,59	Czech Rep.	0,28	Sweden	-0,96
Portugal	1,49	Austria	0,23	Norway	-0,99
Estonia	1,46	Germany	0,00	Malta	-1,10
Croatia	1,28	Belgium	-0,05	Iceland	-1,15

FIGURE 1. Exposure Factors.

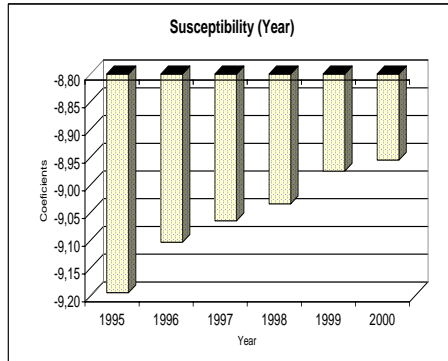and the susceptibility factors, g are presented in Figure 2.

FIGURE 2. Susceptibility Factors.

- And finally

$$\begin{cases} \check{S} = 3.64533E - 25 \\ R^2 \approx 0.9 \end{cases} \qquad (15)$$

These results show a very good adjustment and clearly separate Europe in the following three regions:

- Eastern Europe $(f \geq 1.5)$;

- Balkan Peninsula $(0.5 \leq f < 1.5)$;

- Western Europe $(f < 0.5)$.

With a few exceptions like Portugal.
Moreover a slow but steady decrease of TB incidence is shown across the six years studied.

References

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression.* 2nd edition, Wiley Series in Probability and Statistics.

Mexia, J. T., Pereira, D. and Baeta, J. (1999). L_2 *Environmental Indexes.* Listy Biometryczne - Biometrical Letters Vol. 36, No.2, 137-143.

Surveillance of Tuberculosis in Europe - Euro TB (1995, 1996, 1997, 1998, 1999, 2000). Institut de Veille Sanitaire, Who Collaborating Center for the Surveillance of Tuberculosis in Europe, Royal Netherlands Tuberculosis Association(KNCV).

Series of Studies with a Common Structure: An application to European Economic Integration

Maria M. Oliveira[1], Luis Ramos[2] and João T. Mexia[2]

[1] Departamento de Matemática, Universidade de Évora, Colégio Luis António Verney, Rua Romão Ramalho 59, 7000 Évora, Portugal
[2] Departamento de Matemática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Quinta da Torre, 2825 Monte da Caparica, Portugal

Abstract: An extension of the concept of common structure for a series of studies is presented an applied to European Economic Integration. The significance of Maastrich treaty is clearly seen from that study.

Keywords: European Community; F tests; STATIS method.

1 Introduction

A study will be a matrix triplet constituted by a matrix X of objects x variables, and two diagonal matrices D_n and D_p containing the weights of objects and variables. Escoufier (1973) showed how to obtain geometrical representation of series of k studies when the variables or the objects were the same. In the first case the series will be of first type and, in the second, of the second type. We now extend the concept of common structure of a series of studies given by Lavit (1988).

An application to economic integration of European Union (EU) is presented. The European Community (EC) institutional arrangement was transformed by the Maastricht Treaty originating the European Union. Our results point towards the significance of this institutional transformation.

2 Common Structure

The studies in a series of first type, will be $(\mathbf{X_i}, \mathbf{D_{p_i}}, \mathbf{D_n})$, $i = 1, \cdots, k$, where X_i is the data matrix, while D_{p_i} and D_n are the variables and objects weights matrices. To derive the corresponding geometrical representation Escoufier (1973) obtained the matrix $S = (S_{ij})$ with

$$S_{ij} = Tr(\mathbf{A_i A_j^t}), \; \mathbf{i = 1, \cdots, k}, \; \mathbf{j = 1, \cdots, k} \tag{1}$$

where

$$\mathbf{A_i} = \mathbf{X_i D_{p_i} X_i^t D_n}, \ \mathbf{i} = \mathbf{1}, \cdots, \mathbf{k}. \tag{2}$$

The procedure for series of second type is the same once matrices $\mathbf{A_i}$, $i = 1, \cdots, k$, are replaced by the $\mathbf{B_i} = \mathbf{X_i^t D_{n_i} X_i D_p}, \ \mathbf{i} = \mathbf{1}, \cdots, \mathbf{k}$.

With $(\theta_i, \gamma_\mathbf{i}^\mathbf{k}), \mathbf{i} = \mathbf{1}, \cdots, \mathbf{k}$, the pairs of eigenvalues and corresponding eigenvectors for matrix $\mathbf{S}$, the l-th study was represented by the point whose coordinates $\beta_{l1}, \cdots, \beta_{lk}$ where the l-th components of vectors $\theta_i \gamma_\mathbf{i}^\mathbf{k}$, $i = 1, \cdots, k$, $l = 1, \cdots, k$. Lavit (1988) proposed that a series whose points lie along the first axis had a common structure. It is easily seen that, then

$$\tau_1 = \frac{\theta_1^2}{\sum_{j=1}^{k} \theta_j^2} \approx 1. \tag{3}$$

Inference for such series of studies is presented in Oliveira and Mexia (1998, 1999a, 2004).

We now extend this notation claiming that if $\tau_s = \dfrac{\sum_{j=1}^{s} \theta_j^2}{\sum_{j=1}^{k} \theta_j^2} \approx 1$ the series

has a s-degree structure. The case $s = 2$ is quite interesting since then we have a clear two dimensional image of the set of studies and, find if the studies group themselves into a meaningful pattern.

3   An application to European Economic Integration

We are going to apply our approach to economic integration of EU from 1980 to 2000 since we have not yet enough data to consider the impact of Euro.

For each year we have a study. The objects will be the countries in the EU while the variables will be: Gross Domestic Product, Imports, Exports, Unemployment, Consumption Private, Consumption Public, Industry, Total debit, Total Population and Active Population.

Since the number of countries increased from 10 in 1980 to 15 in 2000 we have a series of second type.

The first two eigenvalues of matrix $\mathbf{S}$ were 389.548 and 74.035. Since $\tau_2 = 4184$ we assumed the existence of a common structure with degree $s = 2$.

In Figure 1 we presente the projections of the points representing the studies in the plane defined by the two first axis.
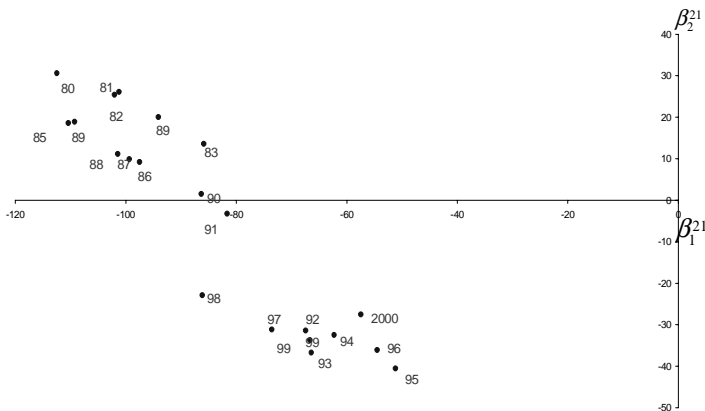
FIGURE 1. Geometrical representation of the studies.

It is a very interesting to point out the clear separation of the years corresponding to EC (80-91) from those corresponding to EU (92-2000). Moreover the points lye along an axis. This led us to center their coordinates and apply principal components. The eigenvalues were 20147.8 and 660.3 so that almost all the information will be carried by the first principal component

$$Y = 0.5999(X_1 + 84.2381) - 0.8000(X_2 + 5.2857). \qquad (4)$$

In Figure 2 we show how the values of that component evolve with time.

FIGURE 2. Evolution of the first principal components.

Despite the linear regression

$$Y = -50.3430 + 4.5766t \tag{5}$$

having an acceptable value for R^2 (0.8005) we must consider that this is due to a linear behavior in the first phase (EC) followed by a second phase (EU) with higher values of Y which seem to oscillate. Thus again we have a separation of the process in two phases:
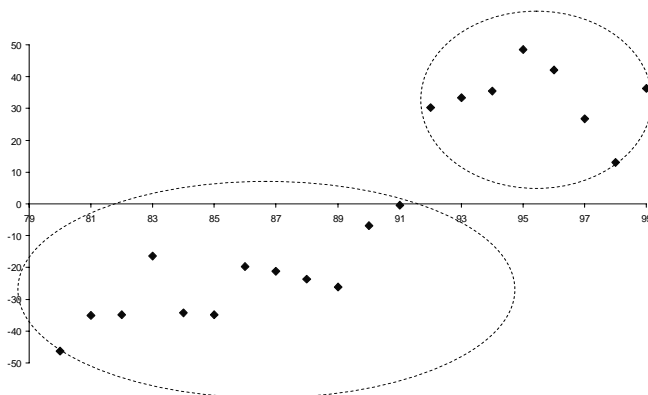
- from 1980 to 1991 when we had EC;

- from 1992 to 2000 when EU was instituted.

As stated above it may be interesting to see, in some years time, if the EURO led to a new phase in the integration.

References

Escoufier Y. (1973). *Le Traitement des Variables Vectorielles.* Biometrics. **29**, *No* **4**, 751-760.

Lavit C. (1988). Analyse Conjointe de Tableaux Quantitatifs. Collection Méthods+Programmes, Masson, Paris 91-262.

Oliveira M.M. and Mexia J.T. (1998). Tests for the rank of Hilbert-Schmidt product matrices. *Advances in Data Science and Classifications.* (Rizzi, Vichi and Bock, Ed.), 619-625. Springer.

Oliveira M.M. and Mexia J.T. (1999a). *F tests for Hypothesis on the Structure Vectors of Series.* Discussiones Mathematicae. Vol. 19, *No* **2**, 345-353.

Oliveira M.M. and Mexia J.T. (2004). *AIDS in Portugal:endemic versus epidemic forecasting scenarios for mortality.* International Journal of Forecasting **20**, 131-135.

Bayesian modelling volatility with mixture of $\alpha-$ stable distributions

Luca Monno[1], Lea Petrella[2] and Andrea Tancredi[2]

[1] Università di Roma3
[2] Università di Roma "La Sapienza"

Abstract: In this paper we propose mixture models for dependent data in time series framework using a Bayesian approach. In particular we build a hidden Markov model with stationary distribution a finite mixtures of α-stable distributions to model time series volatility. Mixtures of α-stable distributions are a very general models that allow for skewness and heavy tails which have as special case the Normal mixtures models. The main problem related with α-stable distributions is the non existence of a close formula for the density function, in order to overcome these difficulies we adopt Markov chain Monte Carlo methods to generate sample for the posterior distribution of the parameters.

1 Introduction

Stable distributions are a rich class of four parameters probability distributions that allow skewness and heavy tails. The non existence of moments of order less than α with $\alpha \in (0, 2]$ and the lack of closed formulas for densities and distribution functions for all but few $\alpha-$stable distributions (Gaussian, Cauchy and Lévy) has been a major drawback to the use of these distributions. Fortunately recently many computer programs have been proposed to handle these distributions and, as a consequence, $\alpha-$stables have been introduced in many different fields as physics, economics, finance and telecomunications. Here we will consider a step forward of modelling time series volatility by constructing a hidden Markov model with stationary distribution a finite mixtures of $\alpha-$stable components. Finite mixtures of distributions have provided a mathematical-based approach to the statistical modelling of a wide variety of phenomena. As any continous distributions can be approximated arbitrarily well by a finite mixture of normal densities with common variance, mixture models provide a convenient semiparametric framework in which to model unknow distributional shapes in particular when attention is focused on tails and skenwess. Mixtures of $\alpha-$stable distributions are a more general model since they have as special case, the mixtures of normal distributions which are the most widley studied finite

mixtures; see for example Richardson and Green (1997). Our aim is to exploit stable mixture models for dependent data in time series framework using a Bayesian approach. In order to overcome the difficulties related with the class of $\alpha-$stable distributions we will adopt Markov chain Monte Carlo (MCMC) methods to generate samples from the full posterior distribution and estimate the parameters following Buckle (1995).

2   Mixture of $\alpha-$stables

A random variable X is said to have a four-parameter stable distribution $S_\alpha(\beta, \gamma, \delta)$ if his characteristic function has the form

$$E(e^{i\vartheta X}) = \left\{ \begin{array}{ll} \exp\left\{-\gamma^\alpha|\vartheta|^\alpha(1 - i\beta(sign\vartheta)\tan\frac{\pi\alpha}{2}) + i\delta\vartheta\right\} & \text{se } \alpha \neq 1, \\ \exp\left\{-\gamma|\vartheta|(1 + i\beta\frac{2}{\pi}(sign\vartheta)\ln|\vartheta|) + i\delta\vartheta\right\} & \text{se } \alpha = 1, \end{array} \right. \quad (1)$$

see Samorodnitsky and Taqqu (1994). The stability parameter α lies in the range $(0, 2]$, and measures the degree of peakedness of the pdf and the heaviness of its tails. When $\alpha = 2$ the stable distributions reduces to a Normal distribution. The skewness parameter $\beta \in [-1, 1]$ measures the departure of the distribution from symmetry, while $\delta \in (-\infty, \infty)$ is the location parameter and $\gamma \in (0, \infty)$ is the scale one. The density function of a finite mixture of $\alpha-$stable distributions would take the form

$$f(x|\Psi) = \sum_{i=1}^{k} \pi_i f(x|\vartheta_i) \quad (2)$$

where the mixing weights are such that $0 \leq \pi_i \leq 1$ and $\sum_{i=1}^{k} \pi_i = 1$; $\vartheta = (\alpha, \beta, \gamma, \delta)$, $\Psi = (\pi_1, \cdots, \pi_k, \vartheta_1, \cdots, \vartheta_k)$ and $f(x|\vartheta)$ is the generic density function of a stable distribution. Some idea of the range of shapes and features provided by mixtures of those distributions are shown in Figure 1(A-D).
Even though mixture models appear to be a simple extension of classical models, they result in complex computational problems when implementing standard estimation principles; in fact due to the assumption that the n observations originate indipendently from the distribution with density (2), the moltiplicative structure of the likelihood function leads to k^n terms. The standard solution to this problem is to use indipendent categorial variables Z taking the values $1, \ldots, k$ with probabilities $\pi_1, \ldots, \pi_k$ defined above, and supposing that the conditional density of X given $Z = i$ is $f(x|\vartheta_i)$. The practical exploitation of the mixture representation from a Bayesian point of view requires the use of Markov chain Monte Carlo simulation, in particular the use of the Metropolis-Hastings within Gibbs Sampler alghorithm will enable us to produce samples from the joint posterior density of the parameters of the mixtures.

In fact, the main problem related within the class of $\alpha-$ stable distributions, i.e. the non existence of a closed formula for the density function, can be overcome introducing an auxiliary variable y such that the stable density is obtained integrating out y from the bivariate density

$$f\left(x,y|\vartheta\right) = \frac{\alpha}{|\alpha - 1|} \exp\left\{-\left|\frac{s}{\tau_{\alpha,\beta}\left(y\right)}\right|^{\alpha/(\alpha-1)}\right\} \left|\frac{s}{\tau_{\alpha,\beta}\left(y\right)}\right|^{\alpha/(\alpha-1)} \frac{1}{|s|} \quad (3)$$

where $(x,y) \in (-\infty, 0) \times (-1/2, l_{\alpha,\beta}) \cup (0, \infty) \times (l_{\alpha,\beta}, 1/2)$ $\tau_{\alpha,\beta} = \frac{\sin(\pi\alpha y + \eta_{\alpha,\beta})}{\cos(\pi y)}$ $\left[\frac{\cos(\pi y)}{\cos(\pi(\alpha-1)y + \eta_{\alpha,\beta})}\right]$, $\eta_{\alpha,\beta} = \beta \min\left(\alpha, 2 - \alpha\right)\pi/2$, $l_{\alpha,\beta} = -\eta_{\alpha,\beta}/\alpha$ and $s = \frac{x-\delta}{\gamma}$. See Buckle (1995)and Casarin (2004) for details.

3   Hidden Markov model with mixture $\alpha-$stable stationary distribution

We shall explore the extent to which mixture of $\alpha-$stable distributions can handle temporally correlated data; specifically, we consider hidden Markov model which have been extensively used to model weakly dependent heterogeneous phenomena, see for example Rydèn *et al.* (1998) and Robert *et. al* (2000). The hidden Markov models extension removes the independence assumption of the mixture models, by considering successive observations from (2) to be correlated through the component k from which they originate. More formally, it is possible to associate to the observations $x_1, \cdots, x_n$ the allocation variables $Z_1, \cdots, Z_n$ having a Markovian structure. Specifically our model will take the form of

$$\sum_{i=1}^{k} \pi_i S_{\alpha_i}(\beta_i, \gamma_i, 0) \quad (4)$$

where the π_i are the components of the stationary vector of the transition matrix of the hidden states $\{Z_1, Z_2, \ldots, Z_t, \ldots\}$ where Z_t is the allocation for the t-th observation, $A = (a_{ij})$, such that $P(Z_{t+1} = j | Z_t = i) = a_{ij}$. Our goal is to model time series which present different regimes of volatility taking advantage of the heterogeneity of the mixture structure. At the same time we can model different frequencies of regime switching by estimating the transition matrix A. To have an idea of the behaviour of the model in Figure 1(E-F) we have considered mixture of a standard normal distribution and a stable distribution $S_{1.5}(0, 1/\sqrt{2}, 0)$ with transition probabilities $P(Z_t = 1 | Z_{t-1} = 0) = 0.1$ $P(Z_t = 1 | Z_{t-1} = 1) = 0.9$ in the top panel and $P(Z_t = 1 | Z_{t-1} = 0) = 0.9$ $P(Z_t = 1 | Z_{t-1} = 1) = 0.1$ in the bottom panel. For the Bayesian inference of the model it is required to derive the form of the full posterior distribution as well as all the complete conditional
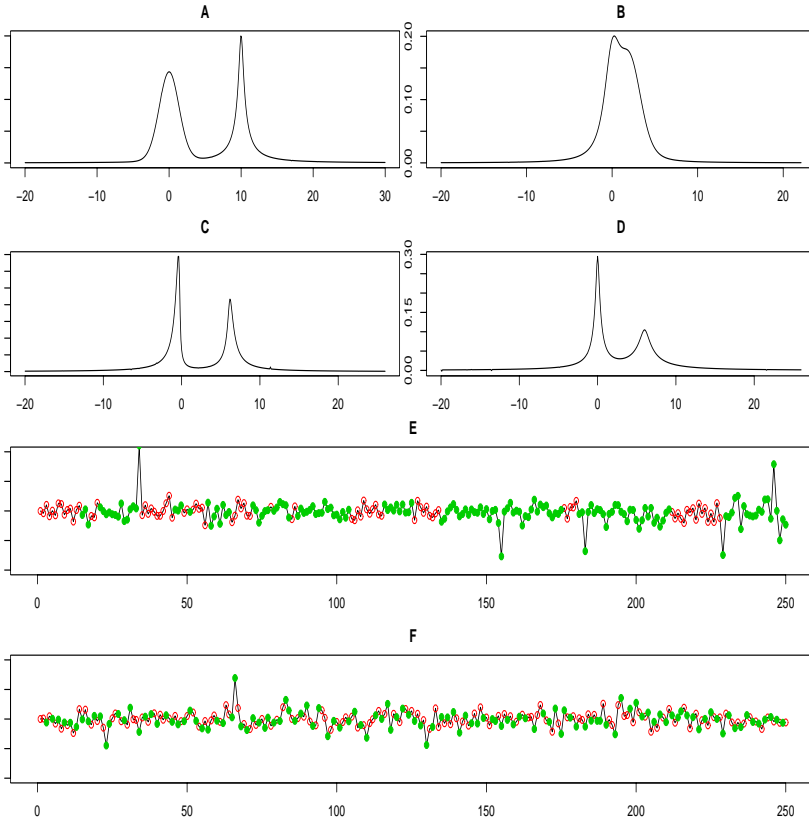
FIGURE 1. A,B,C,D: examples of mixture of α-stable densities. A:$0.5S_{1.9}(0,1,0) + 0.5S_{0.7}(0,1,10)$; B: $0.5S_{1.1}(0,1,0) + 0.5S_{1.7}(0,1,2.1)$; C: $0.5S_{0.7}(-0.8,1,0) + 0.5S_{0.7}(0.3,1,6)$; D: $0.5S_{0.7}(0,0.7,0) + 0.5S_{0.7}(0,2,6)$. E,F: two realizations from a hidden Markov model with mixture α−stable stationary distribution; standard normal realizations=$\circ$; $S_{1.5}(0,1/\sqrt{2},0)$ realizations=$\bullet$. E: $P(Z_t = 1|Z_{t-1} = 0) = 0.1$ and $P(Z_t = 1|Z_{t-1} = 1) = 0.9$; F: $P(Z_t = 1|Z_{t-1} = 0) = 0.9$ and $P(Z_t = 1|Z_{t-1} = 1) = 0.1$.

distributions for the parameters that enable the implementation of the Gibbs Sampler algorithm; the detailed description of one step of our Monte Carlo Markov Chain procedure is as follow:

1. *update transition probability matrix A*: we assume prior independence between the rows of A and a prior distribution for the i-th row $\mathbf{a}_i$ to be a Dirichlet distribution $D(\eta, \cdots, \eta)$. According to that, the conditional distribution of $\mathbf{a}_i$ is $D(\eta + n_{i,1}, \cdots, \eta + n_{i,k})$ where $n_{i,j} =$

$\sum_{t=1}^{n-1} I\{z_t = i, z_{t+1} = j\}$ is the number of jumps from component i to component j;

2. *update the parameter* $\vartheta_i = (\alpha_i, \beta_i, \gamma_i, \delta_i) : i = 1, \ldots k$ generate ϑ_i^l from its complete conditional distribution $\pi(\vartheta_i^l | \vartheta_i^{-l} \mathbf{x}, \mathbf{y}, A, \mathbf{z})$; details on conditional distributions of the specific parameters are shown in Buckle(1995).

3. *update the auxiliar variable* y_t: generate y_t from

$$\pi(y_t | \vartheta, x, A, \mathbf{z}) \propto exp\left\{1 - \left|\frac{s_t}{t_{\alpha,\beta}(y_t)}\right|^{\alpha/(\alpha-1)}\right\} \left|\frac{s_t}{t_{\alpha,\beta}(y_t)}\right|^{\alpha/(\alpha-1)} \quad (5)$$

4. *update the allocations* $\mathbf{Z}$: $Z_1, Z_2, \cdots, Z_n$ are resampled one at a time from $t = 1$ to $t = n$ with conditional probability given by

$$\pi\left(Z_t = i | \vartheta, \mathbf{x}, \mathbf{y}, \mathbf{z}^{-t} A\right) = \frac{a_{z_{t-1},i} f(x_t, y_t | \vartheta_i) a_{i,z_{t+1}}}{\sum_{j=1}^{k} a_{z_{t-1},j} f(x_t, y_t | \vartheta_i) a_{j,z_{t+1}}} \quad (6)$$

when $1 < t < n$, for $t = 1$ the first factor of the numerator is replaced by the stationary probability π_i and for $t = n$ the last factor of the numerator is replaced by 1; here $f(\cdot, \cdot | \vartheta)$ is the joint density (3).

To show how the proposed model can handle volatility we will consider the daily price returns of Abbey National shares already discussed in Buckle (1995).

References

Buckle, D.J. (1995). Bayesian inference for stable distributions. *Journal of the American Statistical Association*, **90**, 605-613.

Casarin, R. (2004). Bayesian Inference for Mixtures of Stable Distributions. CEREMADE Working paper N.0428 University Paris IX.

Richardson, S., Green, P.J. (1997). On Bayesian analysis of mixture with unknown number of components. *Journal of the Royal Statistical Society, Series B*, **59**, 731-792.

Robert, C., Rydèn, T., Titterington, D.M. (2000). Bayesian inference in hidden Markov model through reversible jump Markow Chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, **62**, 57-75.

Rydèn T., Terasvirta T., Asbrink S. (1998). Stylized facts of daily return series and hidden Markov model. *Journal of Applied Econometric*, **13**, 217-244.

Taqqu M.S., Samorodnitsky, G. (1994) *Stable Non-Gaussian Random Processes*. London: Chapman and Hall.

Approximated piecewise linear mixed modelling with random changepoints for longitudinal data analysis

Vito M. R. Muggeo[1]

[1] Dipartimento di Scienze Statistiche e Matematiche 'S. Vianelli' - Università di Palermo, Italy. Email: `vito.muggeo@giustizia.it`

Abstract: In this paper it is discussed how a piecewise linear modelling can be carried out in longitudinal data analysis according to a likelihood based (frequentist) approach. The method, albeit approximated, turns out to be very useful as it allows to obtain both fixed and random effects estimates for each parameter in the model, including changepoints. Data from sieropositive patients are analysed to illustrate the method.

Keywords: changepoint; mixed model; longitudinal data.

1   Introduction and Data

Random effects models are a very useful framework to monitor disease progression in medical studies with repeated measurement design, where several measurements over time are available for each subject. They allow to obtain subject specific estimates of both individual and averaged trajectories, while accounting for heterogeneity, autocorrelation and possible effects of explanatory variables. Although very popular in practice, conventional linear models are not always appropriate, since sometimes the trajectories to be estimated are not linear over the observational follow-up time. This is particulary true in AIDS studies where the decline of some biomarkers' values is not constant but changes at some unknown time-point. This implies that the trend pattern is not simply linear but piecewise linear, exhibiting time-points, the so-called break-points, where it changes rather abruptly: for instance, Lange et al. (1992) and Kiuchi et al. (1995) and references therein, discusse the decline of the number of CD4 T-cell trough such piece-wise modelling in a bayesian perspective.
Difficulties in estimating and testing for such nonstandard models are well-known and are discussed, for instance, in Hall et al. (2003). They also warn about impossibility to fit in a likelihood framework and, as the aforementioned references, perform a fully bayesian analysis to deal with heterogeneity in the changepoints.

Here I propose an approximated method to deal with segmented mixed models in a likelihood-based perspective, generalizing the approach proposed for simple regression models (Muggeo, 2003). To illustrate, I analyse the number of CD4 cell number for $n = 63$ seropositive drug-addicted subjects with 9 measurements each, followed 1989 to 1997 by the Unit of Infection Diseases at University of Catania (Sicilia, Italy). Following Lange et al. (1992) I model the response as square root of the CD4 cell numbers, since such transformation is expected to normalize data.

2   Methodology

Let y_{it} the tth measurement for subject $i = 1, 2, \ldots, n$; the one-breakpoint segmented mixed model is $y_{it} = \beta_{0i} + \beta_{1i}t_i + \beta_{2i}(t_i - \psi_i)_+ + \epsilon_{it}$ where $a_+ = a \times I(a > 0)$ and $I(\cdot)$ is the indicator function. hence for the generic subject i, β_{1i} and $\beta_{1i} + \beta_{2i}$ mean respectively the left and right slopes before and after the changepoint ψ_i and ϵ_{it} is the usual error term with variance σ^2. Assuming only heterogeneity (i.e. no dependence on explanatory variables) in each parameter describing the ith track, the equation becomes

$$y_{it} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_i + (\beta_2 + b_{2i})(t_i - [\psi + p_i])_+ + \epsilon_{it} \quad (1)$$

Here the beta-parameters $(\beta_0, \beta_1, \beta_2, \psi)$ are the fixed effects sometimes said 'population- averaged' (or simply 'population' parameters) and the random effects b, are understood to be able to account for heterogeneity between subjects with respect to corresponding fixed parameters. Therefore for instance, b_{1i} describes how the evolution of the ith subject (before the changepoint) differ from the average β_1, p_i measures how much the ith changepoint deviates from ψ and so on. Typically it is assumed that the random effects are multivariate zero-mean Normal distribution with variance-covariance matrix D, say, $b \sim \mathcal{N}(0, D)$ and independent of the noise ϵ.

Muggeo (2003) shows that the segmented (nonlinear) model has an intrinsically linear form, so generalizing such re-parameterization to a mixed framework leads to linear model:

$$y_{it} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_i + (\beta_2 + b_{2i})U_{it} + (\gamma + g_i)V_{it} + \epsilon_{it} \quad (2)$$

where $U_{it} = (t_i - \hat{\psi}_i^{(0)})_+$ and $V_{it} = -I(t_i > \hat{\psi}_i^{(0)})$ are two variables evaluated at current estimate of breakpoint

$$\hat{\psi}_i^{(0)} = \hat{\psi}_i^{(-1)} + \hat{\gamma}_i / \hat{\beta}_i \quad (3)$$

Here $\hat{\psi}_i^{(-1)}$ is the estimate at the previous step and $\hat{\gamma}_i$ and $\hat{\beta}_i$ are individual (i.e. fixed + random) estimates from model (2). The algorithm starts by putting an initial guess $\hat{\psi}_i = \psi^*$ for every i and goes on by fitting iteratively

model (2) up to convergence that is usually assured if a breakpoint exists. At the final iteration, estimates of the population parameters and predictions for the random effects in the (2) are provided. Fixed and random effects concerning the parameter γ will be not usually noteworthy since such parameter just measures the gap between the two fitted lines (the left and the right slope) at the final estimate of the changepoint. On the other hand, as regard to changepoints, the algorithm also returns the individual estimates $\hat{\psi}_i$ by means of formula (3). These can be used to obtain naïve estimates of the quantities of interest, namely: fixed-effect estimate, $\hat{\psi} = \sum \hat{\psi}_i / n$; zero-mean 'predictions' by the 'residuals' $\hat{p}_i = \hat{\psi}_i - \hat{\psi}$ and relevant standard deviation, $\hat{\sigma}(p) = (\sum \hat{p}_i^2 / n)^{0.5}$.

Of course, when no heterogeneity is assumed in the changepoint, a single fixed estimate can be obtained just by using the fixed estimates of γ and β_2 in the (3).

3   Analysis and Results

Figure 1 left side shows the observed values of the CD4-T cell number (square root) against time for the $n = 63$ aforementioned subjects. Decline in the biomarker's values is rather evident, but the rate does not seem constant as it slows down at approximatively 3 and even 6 years. Based on such empirical evidence a segmented mixed model with two changepoints is fitted; This is

$$y_{it} = \beta_{0i} + \beta_{1i}t_i + \beta_{2i}(t_i - \psi_{1i})_+ + \beta_{3i}(t_i - \psi_{2i})_+ + \epsilon_{it} \qquad (4)$$

Moreover for simplicity the covariance matrix D of the random effects is assumed diagonal, meaning independent random effects. Estimation is performed throughout restricted maximum likelihood.

Table 1 displays parameter estimates for two models with two breakpoints. Model I assumes heterogeneity only in two parameters, the intercept (β_0) and the left slope (β_1). By contrast, in the Model II random effects for each parameter, including changepoints, are accommodated. Estimates for the changepoints (fixed effects and standard deviations) have been obtained trough the aforementioned 'naïve' approach.

Fitted values are plotted in the right side of Figure 1 where some discrepancy between observed and fitted is evident for high values of response at early times. This however might be also due to a misspecified form of matrix D and not depending on the segmented formulation. According results in Table 1, Model II should be preferred meaning that heterogeneity in changepoints and/or difference-in-slope parameters are necessary, although a more parsimonious formulation might be reached. For instance, heterogeneity in the left slope and in the first breakpoint could be ignored as also emphasized in the observed profiles.

Finally for comparison a quadratic model has been also fitted but the fit was worse (AIC=1568.2 on 7 degrees of freedom).
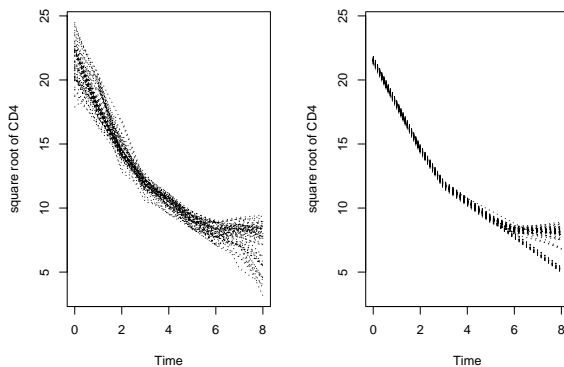
FIGURE 1. Observed (left side) and fitted (right side) individual profiles of CD4 T-cell number (square root) over time (years) for $n = 63$ HIV-positive subjects. The fitted lines come from Model II in Table 1.

TABLE 1. Estimates from two segmented mixed models(see text).

Parameter	Model I		Model II	
	Estimate	$\mid t \mid$ value	Estimate	$\mid t \mid$ value
Fixed Effects				
β_0	21.52	211	21.52	236
β_1	-3.46	45.8	-3.46	51.6
β_2	2.05	19.4	2.13	27.8
β_3	0.98	9.3	1.22	13.3
$\psi_1^{\P}$	2.64	–	2.68	–
$\psi_2^{\P}$	5.38	–	6.83	–
Random Effects (st.dev.)				
b_0	0.263	–	0.220	–
b_1	0.084	–	0.001	–
b_2	–	–	0.040	–
b_3	–	–	0.305	–
$p_1^{\P}$	–	–	0.002	–
$p_2^{\P}$	–	–	3.001	–
σ	0.838		0.752	
AIC (*df*)	1535.6 (*9*)		1409.2 (*13*)	

¶ naïve estimates

4 Conclusions

Here it has been illustrated a method that allows to deal with segmented mixed models in a frequentist context. This is a nontrivial advantage, as previous papers have dealt with the topic only from a bayesian standpoint. Focus has been on estimation and in practice the method seems to work, but several points have to be clarified, including: how calculate standard errors and/or confidence intervals for the changepoints when relevant random effects are included; and how hypothesis testing on heterogeneity in the changepoints may be carry out. Namely how is it possible to test whether all subject have the same changepoint? Likelihood ratio tests comparing the models having the same and different changepoints can be carried out in a straightforward way, but some simulation experiment (here not shown) have emphasized that the null distribution is far from a simple chi-square distribution; in particular such standard tests turn out to be dramatically anti-conservative. However simulations have also shown that the changepoint estimator is asymptotically unbiased and provide predictions reasonably close to true values. Therefore, although further research is needed, the method seems to be a valid frequentist alternative to the bayesian approach.

References

Hall, C.B., Ying, J, Kuo, L., and Lipton, R.B. (2003). Bayesian and profile likelihood change point methods for modeling cognitive function over time. *Computational Statistics & Data Analysis*, **42**, 91-109.

Muggeo, V.M.R. (2003). Estimating regression models with unknown breakpoints. *Statistics in Medicine*, **22**, 3055-3071.

Lange, N., Carlin, B.P., and Gelfand A.E. (1992). Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers (with discussion). *Journal American Statistical Association*, **87**, 615-632.

Kiuchi, A.S., Hartigan, J.A., Holford, T.R., *et al.* (1995). Change points in the series of T4 counts prior to AIDS. *Biometrics*, **51**, 236-248.

A comparison between measure scales for quality evaluation using the Rasch Model

Chiara Zanarotti[1] and Laura Pagani [2]

[1] Institute of Statistics, Catholic Univeristy of Milan, L.go Gemelli, 1, 20122 Milan, Italy

[2] Department of Statistics, University of Udine, Via Treppo, 18, 33100 Udine, Italy

Abstract: The Rasch model is used to compare results obtained using two different rating scales (a four point and a five point rating scale) in evaluation of quality of a public service (university courses). Through the Rating Scale Rasch Model the performance of the two scales are investigated and questionnaire calibration is performed to obtain a more coherent measure tool.

Keywords: Ordinal Data; Rasch Models; Customer Satisfaction; Educational Services.

1 Introduction

In recent years there has been a considerable increase in the practice of collecting customers opinions about different services (private as well as public ones) in order to measure the quality of those services. In collecting opinions researchers typically use questionnaires formed by a few questions or items regarding different aspects of the service and customers compile these questionnaires. Usually the customer can choose the response to each item among a set of given categories or scores. For example, answers can be ordinal categories that vary from *very dissatisfied* (or very insufficient, or strongly disagree) to *very satisfied* (or very good, or strongly agree). The answers of any customer to each item depend not only on service quality, but also on personal characteristics and on the measure tool used for gathering information. Even if quality is the same, different customers can give it different evaluations because of personal characteristics. Responses also are influenced by the choice of items included in questionnaires, or by their lexical formulation, or by the response categories. A very powerful model able to treat this kind of data is the Rasch Model (Rasch, 1960). Introduced in psychometric field, the Rasch model has had increasing success in other applied fields both for its flexibility and simplicity, and also because of the robustness of its measures. As pointed out by many authors (see, for example Molenaar and Fisher, 1995; Bond and Fox, 2001; Beltyukova and Fox, 2002; Tesio, 2003), this model represents a very appealing way to obtain

universal, objective measures in the social sciences. The Rasch model, in fact, is a latent structure model by means of which it is possible to derive continuous measures from total scores obtained by a set of subjects on a set of items. One of the more interesting aspect of the Rasch model is that it is a falsifiable model, in the sense that it is possible to detect items (or subjects) that are incoherent and have to be deleted from the questionnaire. The aim of this paper is to study the effects of two different rating scales on quality measure using the Rasch model. The model considered is the polytomous extension of the original (dichotomous) Rasch model. In particular, the so-called Rating Scale Models will be used (see, for example, Bond and Fox, 2001, ch. 6), which is a suitable model when response categories are rating scale type with the same number of categories for all items of questionnaire. The basic assumption of the Rasch model is that the response of any subject to each item depends on two parameters: a person parameter, reflecting personal subjective characteristics, and an item parameter, that measures each item quality, i.e. the item position along an interval scale reflecting its quality level. Suppose that J items have been administered to I persons and that each item has K ordinate response categories. Let X_{ij} be response of person i $(i = 1, \ldots, I)$ to item j $(j = 1, \ldots, J)$. The *Rating Scale Rasch Model* (RSRM) assumes that probability that person i $(i = 1, \ldots, I)$ chooses response k $(k = 1, \ldots, K)$ for item j $(j = 1, \ldots, J)$ instead of response $k - 1$, is:

$$P(X_{ij} = k | \beta_i, \delta_j, \tau_k) = \frac{\exp\{\beta_i - \delta_j - \tau_k\}}{1 + \exp\{\beta_i - \delta_j - \tau_k\}} \tag{1}$$

In RSRM the function that links individual response probability to parameters is the logistic transformation. Probabilities depend on three sets of parameters: person parameters β_i $(i = 1, \ldots, I)$, item parameters δ_j $(j = 1, \ldots, J)$ and threshold parameters τ_k $(k = 1, \ldots, K-1)$. The use of Rasch models in quality evaluation of a service, as already pointed out by others authors (for example, Bertoli-Barsotti and Franzoni, 2001), implies the following meaning for the parameters:

- person parameters (also called *person location*) measure individual satisfaction and reflect all personal characteristics that can influence satisfaction. High values of person parameter means highly satisfied persons, while low values means the reverse;

- item parameters (also called *item location*) measure quality related to each item. High values of item parameter means service aspect with low quality, while low values means the reverse; so it is possible to measure and order items from the one showing best quality to the one showing worst quality;

- threshold parameters measure the difficulty to endorse each response category over the previous one. In RSRM all items are supposed to

have the same number of categories (K) and distances between adjacent categories are supposed to be the same for every item. Each parameter τ_k represents the cut-off point between category k and category $k + 1$.

2 Analysis of quality of university courses using RSRM: scale impact

The aim of this paper is to measure the quality level of the teaching service using two different sets of rating scale: a four point scale (with labels 1=*not at all satisfied*, 2=*dissatisfied*, 3=*satisfied*, 4=*very satisfied*) and a five point scale (with labels 1=*not at all satisfied*, 2=*dissatisfied*, 3=*almost satisfied*, 4=*satisfied*, 5=*very satisfied*). The analysis is performed using the RSRM (see section 1). Data considered are responses given for two kinds of questionnaires (360 for the four point and 441 for the five point scale) randomly administrated during year 2000 to students of the Faculty of Economics, University of Udine, Italy. We focus our analysis on the comparison of the behavior of Item Location Parameters in the two scales (without consider the same analysis of the Person Location Parameters) because our interest lies with the calibration of the questionnaire. Data collected, summarized in Table 1, have been analyzed with RUMM 2010 (RUMM Laboratory Pty Ltd), a standard software for the Rasch analysis.

TABLE 1. Percentage frequencies of item responses.

Item Code	Item Label	Four-point Scale				Five-point Scale				
		1	2	3	4	1	2	3	4	5
d13	Meets course objectives	4	17	56	23	2	10	29	42	17
d14	Indicates how to prepare the course	5	33	48	14	3	20	37	29	11
d15	Develops the course systematically	3	16	63	18	4	10	30	39	17
d16	Outlines the major points clearly	5	18	57	20	4	10	29	39	18
d17	Links to other subjects	8	43	42	8	5	23	41	27	5
d18	Provides examples and case studies	3	18	58	21	2	9	30	39	20
d19	Explains clearly	13	24	40	23	11	18	24	29	17
d20	Motivates the students	8	36	38	18	6	20	31	31	12
d21	Gives deeper understanding of topics	2	17	59	22	2	9	33	42	15
d22	Is punctual	6	10	44	40	4	5	19	35	37
d23	Is accessible to students	1	8	53	38	1	3	22	45	28
d24	Has a genuine interest in students	2	12	46	40	3	3	19	44	31
d25	Quality of text books and notes	5	21	64	10	4	10	42	38	6
d26	Effectiveness of other materials	5	27	58	11	3	16	38	37	7
d27	Quantity of time for exercises	4	28	59	9	5	16	37	37	5
d28	Utility of exercises, laboratory, etc.	7	17	57	19	4	12	34	38	12
d29	Links between lectures and exercises	4	22	61	13	4	15	37	37	8
d30	Satisfaction level of exercises	6	23	58	14	5	15	35	38	7
d31	Global satisfaction	3	19	62	15	4	12	30	42	12

In Table 2, giving the estimates of the item location parameters (ILP) ordered from the lowest to the highest, it can be observed that the items with best quality (with negative location value) and items with lowest quality (with positive location value) are the same, apart from their order.

TABLE 2. Estimated item location parameters for the two scales.

	Four-point Scale					Five-point Scale			
IC	ILP	SE	Chi Sq	Prob	IC	ILP	SE	Chi Sq	Prob
d23	-1.21	0.10	3.27	0.95	d24	-1.05	0.07	9.37	0.40
d24	-1.08	0.10	16.39	0.06	d22	-0.99	0.07	13.82	0.13
d22	-0.85	0.10	19.23	0.02	d23	-0.98	0.07	7.09	0.63
d21	-0.34	0.09	7.18	0.62	d18	-0.34	0.07	10.77	0.29
d13	-0.27	0.09	8.14	0.52	d13	-0.25	0.07	7.38	0.60
d18	-0.22	0.09	4.23	0.90	d21	-0.22	0.07	14.91	0.09
d15	-0.21	0.09	12.66	0.18	d16	-0.19	0.07	22.94	0.01
d16	-0.11	0.09	9.42	0.40	d15	-0.15	0.07	12.84	0.17
d31	-0.02	0.09	27.38	0.00	d31	0.00	0.07	26.12	0.00
d28	0.02	0.09	27.48	0.00	d28	0.13	0.07	23.94	0.00
d29	0.17	0.09	9.47	0.40	d25	0.26	0.07	84.47	0.00
d30	0.27	0.09	7.42	0.59	d29	0.33	0.07	6.97	0.64
d25	0.28	0.09	50.45	0.00	d26	0.35	0.07	7.30	0.61
d26	0.42	0.09	9.90	0.36	d30	0.40	0.07	16.06	0.07
d27	0.44	0.09	21.14	0.01	d14	0.44	0.06	5.09	0.83
d19	0.50	0.09	59.73	0.00	d19	0.47	0.06	85.03	0.00
d14	0.55	0.09	14.90	0.09	d27	0.48	0.06	28.17	0.00
d20	0.61	0.09	36.50	0.00	d20	0.49	0.06	28.16	0.00
d17	1.06	0.08	17.65	0.04	d17	0.81	0.06	20.76	0.01

The two central items (d31 e d28) are the same for the two scales. This means that items' order is scale dependent.

TABLE 3. Estimated threshold parameters for both scales.

Four-point Scale			Five-point Scale			
1 to 2	2 to 3	3 to 4	1 to 2	2 to 3	3 to 4	4 to 5
-2.125	-0.463	2.588	-2.224	-1.047	0.629	2.642

In Table 3 the estimated thresholds are reported and one observes that distances between adjacent thresholds are very different. In the four-point scale while distance between threshold one and threshold two is 1.66, the distance between threshold two and threshold three is 3.05. Also in the five point scale differences are not the same, but change significantly. This suggests avoiding the use of natural numbers, like $1, 2, \ldots$, to quantify categories in quantitative analysis. In the first two columns of Table 4 there is a summary of the models fit including all items.
The results show that the global Chi-square, for both scales, is not statistically significant.
 Looking again at Table 2 is possible to investigate which items, having associated low values of p-value, give the major contribution to the global Chi-square values. For the four point scale there are five items that don't fit, while they are seven for the five point scale.
The RSRM is estimated again, for both scales, after exclusion of those items from the models. The fitting tests obtained for the resulting models

TABLE 4. Models fitting for both scales.

	All items		After deleting some items	
Scale	Chi-square test	D. F.	Chi-square test	D.F.
Four-point	362.52	171	145.79	126
Five-point	431.17	171	122.60	90

are reported in the last two columns of Table 4. The Chi-square test is statistically significant for the four point scale and is insignificant for four the five point scale.

3   Conclusions

In conclusion, it is possible to hypothesize that results are scale dependent. A major result is that considering all categories as equidistant is misleading. From the goodness-of-fit statistics is possible to note that the five point scale seems to be more problematic than the four point scale.

References

Beltyukova S.A., Fox C.M. (2002). Student Satisfaction as a Measure of Student Development: Towards a Universal Metric. *Journal of College Student Development, Vol.43*, **2**, 1-12. Oxford: Clarendon Press.

Bertoli-Barsotti L. and Franzoni S. (2001). *Analisi della soddisfazione del paziente in una struttura sanitaria: un caso di studio*. Universit Cattolica, Milano, Serie E.P., **104**, 1-17.

Bond T.G. and Fox C.M. (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Lawrence Erlabaum Associates Publishers, Mahawah, New Jersey.

Fischer, G.H., Molenaar, I.W. (1995). *Rasch Models. Foundations, Recent Developments, and Applications*. Spriger-Verlag, New York.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research, Copenhagen.

Tesio, L. (2003). Measuring Behaviours and Perceptions: Rasch Analysis as a tool for Rehabilitation Research. *Journal of Rehabil. Med.*, **35**, 105-115.

On probabilities of avalanches triggered by alpine skiers. An application of models for counts with extra zeros

Christian Pfeifer[1] and Verena Rothart[1]

[1] Institut fr Statistik, Universitt Innsbruck, Universittsstrae 15, A–6020 Innsbruck. e-mail: `christian.pfeifer@uibk.ac.at`

1   Introduction and problem

In Austria most fatal snow avalanche accidents are caused by skiers or snowboarders. For example in winter 2001/02 79 avalanche accidents (17 fatalities) are reported. 16 from 17 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 (1997) suggested a quantitative method to estimate the risk of avalanche events. Assumig that variates

- danger levels from the local avalanche information service (1=low to 5=very high)

- incline of the slope (3 classes from 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". As a consequence of this several other strategies were developed in order to estimate avalanche danger when backcountry skiing (Plattner, 2001). But as we showed in Pfeifer and Rothart (2002) Munter's quantity cannot be seen as 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. At least it is necessary to include some information on frequencies of skiers on slopes under specific conditions.

2   First statistical model

In Rothart and Pfeifer (2003) we proposed a statistical model on the counts y_i of avalanche events in each class of incline and aspect for days i with avalanche reports from the Tyrolean avalanche information service (Lawinenwarndienst Tirol).

$$\log(y_i) = \texttt{LWS} + \texttt{NEIG} + \texttt{EXPOS} + \texttt{WOENDE} + \texttt{TOURV}$$

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 `WOENDE` 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 2000-2002 reported by the Tyrolean avalanche information service (497 days of observation). Because avalanche accidents are expected to be rather rare this simple Poisson model shows strong underdispersion (residual df = 2975, residual deviance = 645.43). In the following we employ 2 models to overcome this misspecification:

3   Models for counts with extra zeros

Zero inflated Poisson models (**ZIP**) assume observations y_i to be from a mixture of a Bernoulli and Poisson distribution:

$$\mathrm{P}(y_i) = \begin{cases} 1 - p + p\exp(-\lambda) & : \quad y_i = 0 \\ \frac{p\exp(-\lambda)\lambda^{y_i}}{y_i!} & : \quad y_i > 0 \end{cases}$$

The observations of zero altered Poisson models (**ZAP**) are assumed to come from a mixture that is zero with probability one in the first component and a truncated Poisson in the second component:

$$\mathrm{P}(y_i) = \begin{cases} 1 - p & : \quad y_i = 0 \\ \frac{p\exp(-\lambda)\lambda^{y_i}}{(1-exp(-\lambda))y_i!} & : \quad y_i > 0 \end{cases}$$

In the first case the response variate can be seen to be dependent on an unobserved indicator z which is equal to zero if y_i is a structural zero and equal to 1 if y_i is from Poisson distribution. One could say that it is inappropriate to distinguish between structural zeros of the Bernoulli process and sampling zeros of the Poisson process. The second approach, however, does not make a difference between two states of zeros.

In order to define the covariates effects on the observations we use the link functions of the logistic and the loglinear model:

$$\log(\lambda) = \mathbf{B}\beta \qquad \qquad \mathrm{logit}(\mathbf{p}) = \mathbf{G}\gamma$$

TABLE 1. shows results (parameter estimates, standard errors and log-likelihood) for the Poisson, the ZIP(τ) and the ZAP model (The ZIP model in the unrelated case did not show reliable results):

	Poisson		ZIP(τ)		ZAP			
	β	se	β	se	β	se	γ	se
ICPT	-7.025	0.584	-5.426	1.278	-4.734	3.617	-7,228	0.642
LWS	0.937	0.165	0.805	0.242	1.491	1.075	0.912	0.178
NEIG	0.795	0.136	0.678	0.193	0.031	0.619	0.833	0.147
EXPOS	-0.541	0.200	-0.464	0.203	-0.188	0.731	-0.578	0.216
WOENDE	0.323	0.199	0.292	0.186	-0.363	0.846	0.401	0.215
TOURV1	-0.314	0.256	-0.248	0.245	-1.765	0.747	-0.123	0.291
TOURV2	-1.090	0.343	-0.928	0.389	-2.431	1.439	-0.937	0.382
τ			0.302	0.363				
loglik	-417.10		-414.11		-410.55			

If the covariate matrices $\mathbf{B}$, $\mathbf{G}$ and the parameter vetors β, γ are independent, λ and γ are assumed to be unrelated. In order to reduce the number of parameters it is recommended to define a relationship between λ and γ as follows:

$$\text{logit}(\mathbf{p}) = \tau \mathbf{B} \beta$$

The linear predictor of the logistic part depends on the linear predictor of the loglinear part $\mathbf{B}\beta$ and a real valued shape parameter τ. Technical details to these models are given in Lambert (1992) and Welsh et al. (1996).

4   Calculation and results

We fitted ZIP and ZAP models for the same parameter vector as in the Poisson case. Maximum likelihood estimates of the parameters were computed with a quasi-Newton algorithm (implemented in the `Splus` function `nlminb`). In the case of ZAP models we used the `Splus` function `ezp` provided by Heather M. Podlich in the `Splus`-library `extraz45` (www.maths.uq. edu.au/ hmp/extraz.html).

5   Conclusion

Using models for counts with extra zeros seems to increase the goodness of fit of the Poisson model. Predicted probabilities are slightly lower than in the Poisson case. If we pay our attention to predicted probabilities there is almost no difference between the ZIP(τ) and the ZAP model.
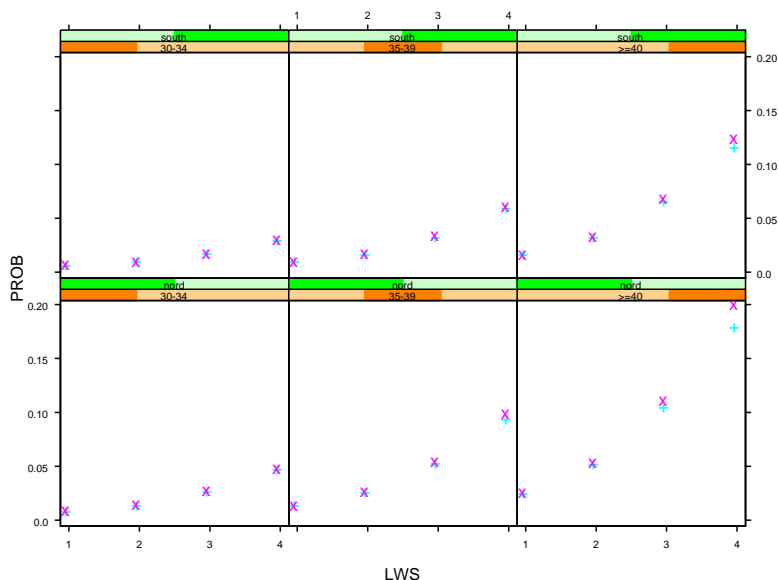
FIGURE 1. shows predicted probabilities of avalanche counts larger or equal than 1 dependent on danger level LWS, incline of slope NEIG and aspect of slope EXPOS for the Poisson (x) and the ZIP(τ) model (+). There is almost no difference between predicted probabilities of the ZIP(τ) and the ZAP model.

References

Lambert, D. (1992). Zero-Inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1-14.

Munter, W. (1997). 3x3 Lawinen, Pohl & Schellhammer, Garmisch-Partenkirchen.

Pfeifer, C., Rothart, V. (2002). Die Reduktionsmethode zur Beurteilung der Lawinengefahr fr Schitourengeher aus statistischer Sicht; Jahrbuch der Kuratoriums fr alpine Sicherheit 2002, Innsbruck.

Plattner, P. (2001). Werner Munter's Tafelrunde; Berg & Steigen 4/01, Innsbruck.

Rothart, V., Pfeifer, C. (2003). Neuere Methoden zur Beurteilung der Lawinengefahr fr Schitourengeher aus statistischer Sicht. Ein erstes statistisches Modell mit Informationen von Begehungsfrequenzen; presentation at sterreichische Statistiktage 2003 31.10.2003 Vienna

Welsh, A.H. et al (1996). Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling*, **88**, 297–308.

Fieller's method for mixed models

Birgitte B.Rønn[1]

[1] Biostatistics, Novo Nordisk A/S, Novo Nordisk Alle, 2880 Bagsværd, Denmark,
Email: BBR@novonordisk.com

Abstract: In complete, balanced dose-response trials with independent observations, the dose-ratio between different treatment groups can be estimated and exact confidence limits can be found by the Fieller method. We have used the same simple approach on the general parallel line model and found Fieller-type confidence limits derived from approximate distributions. A simulation study show that in situations similar to the dose-linearity trial, the approximate distributions seem to fit reasonably for samples, as small as 5 subjects per group. Furthermore confidence regions based on the approximate distribution results in far more sound conclusions than regions obtained by the delta method, relying on asymptotic results, when the dose-ratio is truly a non-linear function of the parameter estimates.

Keywords: Pharmacokinetics; Dose-response trials; Mixed models.

1 Introduction

In pharmacokinetics it is often of interest to compare the dose-concentration relations of two drugs or of the same drug administered by different administration routes. As a part of this comparison the relative bio-availability might be of interest, that is, a comparison of administered doses resulting in the same measurable concentration of drug in the blood. In Figure 1 individual log(dose)-log(concentration) profiles are shown from a dose-linearity trial, for two different administration routes of insulin A and B. The trial was a five period cross over trial in 21 type 1 diabetic subjects. Each subject was randomized to five of seven possible treatments (three insulin doses administered by A and four insulin doses administered by B). The individual log(dose)- log(concentration) relations appear to be linear for both administration routes and further the linear relations seem parallel. Within a parallel line model, with log-transformed concentration as response variable and log-transformed dose as co-variate, the ratio between doses that result in same measured blood concentrations corresponds to the horizontal distance between the estimated lines. The estimate is then a non-linear function of the slope and intercepts, and the confidence limits can be found by approximate methods e.g. the delta method. For balanced designs with homogeneous variance, exact confidence limits for the estimated bio-availability has been developed by (Fieller, 1940). Extensions to cross-over designs can be found in (Finney, 1978). In a multidimensional setting, where dose ratios resulting in equal response with respect to several properties are of interest, the exact method of Fieller in a generalized form, can be
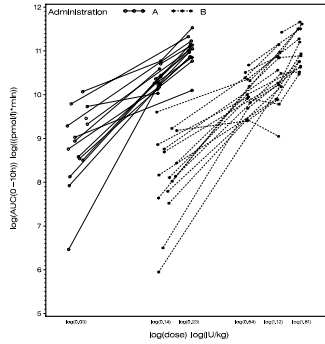
FIGURE 1. Individual log(dose)-log(concentration) relations.

applied, see (Vølund, 1980). The above mentioned clinical trial was meant to be balanced, but problems occurred in a few number of experiments and the data available for analysis did not reflect the planned, balanced design. Pharmacodynamic quantities were also measured during the trial and hence the procedure at each visit were rather demanding for the subjects and not all completed the visits as planned. Therefore, it can be discussed whether it is reasonable simply to disregard an amount of information obtained for the non-completing subjects in order to achieve balanced data. Furthermore, it should be noted that administration route B seems to result in more fluctuating and less stable log(dose)-log(concentration) relations than B, reflecting heterogeneous within subject variances in the groups. Since the trial was a cross-over study also between-subject variability should be accounted for within the model. Finally, it is seen that for both administration routes measurements corresponding to the lowest dose are much more variable than measurements corresponding to the higher dose levels. A unbalanced, mixed model with a complex covariance structure should be fitted to the data and the bio-availability with confidence limits should be estimated within this model.

2   The parallel line model

The concept of relative bio-availability only makes sense with parallel log(dose) - concentration relations between treatment groups, since the definition as ratio between doses, resulting in the same concentration (response) then becomes constant. In pharmacokinetics it is reasonable to assume dose-linearity, since the amount of drug administrated is usually proportional to the amount of drug measured in the blood. Hence, the log(dose)-log(concentration) relation can be modelled by a parallel line model and concentrations are often log-normally distributed and normal theory can be applied. The general parallel line model is an ordinary mixed model,

$$\mathbf{Y} \;=\; \mathbf{X}^t\beta + \mathbf{Z}^t\gamma + \varepsilon, \tag{1}$$

where $\mathbf{Y}$ is the vector of measurements, $\mathbf{X}$ is the design matrix for the fixed part, β is the fixed parameter, $\mathbf{Z}$ is the design matrix for the random part, γ is the random effects vector and ε is the vector of random errors. The random vectors γ and ε are assumed to be mutually independent and normally distributed with variance matrices Ω and Σ. Now assume that the contrast between the two treatment groups of interest is given by β_0 and the common slope in log(dose) is given by β_1. The relative bio-availability μ is then given by $\mu = \frac{\beta_0}{\beta_1}$ and estimated by $\hat{\mu} = \frac{\hat{\beta}_0}{\hat{\beta}_1}$. The estimator is a non-linear function of the parameter vector. Within the general mixed model set-up we follow the approach of Fieller and consider the hypothesis,

$$\mathcal{H}_0 \quad : \quad \hat{\beta}_0 - \mu\hat{\beta}_1 = 0 \tag{2}$$
$$: \quad L^t\beta = 0, \tag{3}$$

with $L = (1, -\mu)^t$. After normalization with the standard deviation a t-statistic for the hypothesis can be calculated, $\hat{t} = \frac{L\beta}{\sqrt{L\hat{W}L^t}}$. The $\hat{t}$ is approximately t-distributed with degrees of freedom that needs to be estimated from the data, see e.g. (Verbeke and Molenberghs 2000). The t-distribution leads to following equation, determining the acceptance area at significance level α,

$$(\hat{\beta}_0 - \mu\hat{\beta}_1)^2 \quad \leq \quad t_{\hat{df},\frac{\alpha}{2}}^2 L\hat{W}L^t. \tag{4}$$

The equality corresponds to a second order equation for the confidence limits with solutions

$$\mu_{\text{lower}}, \mu_{\text{upper}} \quad = \quad \frac{\hat{\beta}_0\hat{\beta}_1 - t_{\hat{df},\frac{\alpha}{2}}^2 \hat{W}_{01} \pm t_{\hat{df},\frac{\alpha}{2}}\sqrt{A}}{\hat{\beta}_1^2 - t_{\hat{df},\frac{\alpha}{2}}^2 \hat{W}_{11}} \tag{5}$$

where $A = \hat{\beta}_1^2\hat{W}_{00} + \hat{\beta}_0^2\hat{W}_{11} - 2\hat{\beta}_0\hat{\beta}_1\hat{W}_{01} + (\hat{W}_{01}^2 - \hat{W}_{00}\hat{W}_{11})t_{\hat{df},\frac{\alpha}{2}}^2$. The confidence limits given by the equation are borders of an acceptance area and not defined from the distribution of the estimator. The limits are found as roots in a second order equation and then no real valued solution need to exist and the estimated relative bio-availability need not to be included in the confidence interval. Further, the confidence limits are based on approximate distribution results, but for small samples this approach might result in more reasonable confidence limits, than the delta method, which is based on asymptotical normality combined with a crude first order Taylor expansion.

3   Simulation study

A simulation study was done to investigate the properties of the confidence limits. The simulations were inspired by the dose-linearity trial described

above, with regard to design and covariance structure. The parallel line model was assumed,

$$Y_{ij} = \alpha_{treat(i,j)} + \beta log(dose_{i,j}) + V_{ij}, \tag{6}$$

with treatment dependent intercept and common slope and several different covariance structures, $V(Y_i)$, are simulated. The covariance structure corresponding to the dose-linearity trial, where the heterogenous variability between administration groups and between the lowest dose level and the remaining dose levels are modelled by four measurement error variances, corresponding to administration route A and lowest dose level, administration route A and higher dose levels, administration route B and low dose level and administration route B and higher dose levels. The covariance matrix looks as follows,

$$V(Y_i) = \omega^2 \cdot J + \sigma^2_{treat(j),low(j)} \cdot I, \tag{7}$$

where J is a n by n matrix of ones, I is the n dimensional identity matrix, $treat(i)$ denotes the treatment corresponding to visit j and $low(j)$ indicates whether the dose given at visit j was low or not. Simulations of 1000 samples of 5 or 15 subjects are made and in order to obtain unbalanced designs censoring are introduced of 0, 20 or 50% of the measurements. Simpler structures of the covariance are also simulated, namely independence, a split-plot model and a split-plot model with treatment dependent within subject variance,

$$V(Y_i) = \sigma^2 \cdot I, \tag{8}$$

$$(Y_i) = \omega^2 \cdot J + \sigma^2 \cdot I, \tag{9}$$

$$V(Y_i) = \omega^2 \cdot J + \sigma^2_{treat(j)} \cdot I. \tag{10}$$

Simulations are made with a ratio between the effect of treatments of 40% ($\alpha_A - \alpha_B = 0.4$) and with slopes equal to 1 and 1.8 ($\beta = 1 or 1.8$) corresponding to a linear and a non-linear bio-availability respectively. For all models, the 1000 simulated samples result in Fieller-type confidence limits and limits found by the delta-method.

Simulations from the model with $\beta = 1.8$, that is when the bio-availability is truly a non-linear function of the parameter estimates, the Fieller-type estimates seems to contain the true bio-availability in close to 95% of the samples, whereas the delta-type intervals contain the true parameter too often. However, for samples of 5 subjects with censoring of 50% of the measurements, the models with complex covariance structure fails to fit the data for many simulated samples.

Simulations from the model with $\beta = 1$, where the bio-availability is actually a linear function of the estimated parameters, both the Fieller-type and the delta-type confidence intervals contain the true parameter in close to 95% of the samples.

4   Conclusion

Fiellers exact method for calculating confidence intervals as an acceptance area, is generalized to an approximate method in the general mixed model set-up. The simulation study indicates that the method is better than the conservative delta method in the truly non-linear case, and equally good in the linear case. However in a number of the simulated samples the Fieller confidence interval could not be calculated to contain the estimated bio-availability, but since (Gleser and Hwang, 1987) showed that confidence intervals for rates of regression coefficients will have length with infinite expectation, some problems could be expected. Further the Fieller method is known to be sensitive to small slopes, but in pharmacodynamic studies, where the relative potency of a drug is of interest and where the dose-response relation is seldom proportional, the present results indicates that conclusions based on Fieller-type confidence intervals lead to more reliable conclusions, than regions obtained by the delta method, relying on asymptotic results.

References

Fieller, E (1940). The biologicl standardization of insulin. *J. Roy. Statist. Soc., Suppl,* 7:1-64.

Finney, D.J. (1978). *Statistical Methods in Biological Assay*. London and High Wycombe: Charles Griffin and Company LTD.

Gleser, L.J. and Hwang, J.T. (1987). The nonexistence of $100(1-\alpha)\%$ confidence sets of finite expected diameters in errors-in-variables and related models. *Ann. Statist.,* 4:1351-1362.

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

Vølund, A (1980). Multivariate Bioassay. In: *Biometrics*. 36:225-236.

Predictive model selection criteria for logistic regression

Paolo Vidoni[1]

[1] Department of Statistics, University of Udine, via Treppo 18, I-33100 Udine (Italy); e-mail: vidoni@dss.uniud.it

Abstract: This paper regards an information criterion for model selection proposed by Vidoni (2003). This criterion, based on a predictive density which improves the estimative one, suitably generalizes the Akaike Information Criterion (Akaike, 1973). The theoretical issues, behind this new criterion, are briefly reviewed and an application concerning variable selection under logistic regression models is presented.

Keywords: AIC, model selection, logistic regression, predictive density.

1 Introduction

Let us consider the sample $Y = (Y_1, \ldots, Y_n)$, with $Y_1, \ldots, Y_n$ independent random variables, and a parametric statistical model, specified by the family of probability density functions $\{f(y; \omega), \ \omega \in \Omega \subseteq \mathbf{R}^d\}$, with respect to a common dominating measure, where ω is an unknown d-dimensional parameter, $d \geq 1$. Since there could be several plausible parametric statistical models for Y, we are interested in defining a convenient procedure for model selection. In particular, we aim to choose the model which offers the most satisfactory predictive explanation to the observed sample $y = (y_1, \ldots, y_n)$.

The well-known Akaike Information Criterion (Akaike, 1973), abbreviated as AIC, is defined as a first-order unbiased estimator for a target quantity related to the expected Kullback-Liebler information between the true unknown density of a potential future observation and the corresponding estimative predictive density. More precisely, if the future random vector Z is an independent copy of Y, the theoretical target quantity is

$$\hat{\eta}(g, f) = E_Y[E_Z\{\log f(Z; \hat{\omega})\}]. \tag{1}$$

Hereafter, the expectations are with respect to the true unknown distribution. Indeed, $g(\cdot)$ is the true unknown density of Y and Z and $f(z; \hat{\omega})$ is the estimative or *plug-in* predictive density for Z, under the assumed parametric statistical model, based on the maximum likelihood estimator $\hat{\omega} = \hat{\omega}(Y)$. The AIC selects the model maximizing $\Psi_{AIC}(Y; f) = \log f(Y; \hat{\omega}) - d$, which

is a first-order unbiased estimator for (1), provided that the model under consideration is "true" or it is a good approximation to the truth. An extension of the AIC, not relying on this strong assumption, is the Takeuchi's information criterion (TIC) for model selection (Shibata, 1989). A further generalization of the AIC and the TIC is proposed by Pan (2001) and it is based on the quasi-likelihood approach.

However, both the AIC and the TIC, and their potential extensions, are based on the estimative predictive density, which may be a rather inaccurate estimator for the true density of Z. For this reason, Vidoni (2003) proposed a new information criterion, based on an improved predictive density, as reviewed in the next section.

2   Improved information criterion for model selection

We shall assume the repeated index convention, so that summation is intended over indices that appear more than once in a single term. Following Corcuera and Giummolè (2000), we consider the predictive density $\tilde{f}(z; \hat{\omega})$, which gives the optimal improvement over the estimative one, as estimator of the true density $g(z)$, in terms of average Kullback-Liebler divergence. Namely,

$$\tilde{f}(z; \hat{\omega}) = f(z; \hat{\omega}) \Big[1 + \frac{1}{2} \Big\{ \ell_{rs}(\hat{\omega}; z) + \ell_r(\hat{\omega}; z)\ell_s(\hat{\omega}; z) - \lambda_{rst}(\hat{\omega})\ell_t(\hat{\omega}; z) \Big\} \sigma_{rs} \Big],$$

where, $\ell_r(\hat{\omega}; z)$ and $\ell_{rs}(\hat{\omega}; z)$, $r, s = 1, \ldots, d$, are the first and the second partial derivatives of $\log f(z; \omega)$ with respect to the components of $\omega = (\omega_1, \ldots, \omega_d)$, evaluated at $\omega = \hat{\omega}$, and $\lambda_{rst}(\hat{\omega})$ is a suitable coefficient specified by Corcuera and Giummolè (2000). Furthermore, $\sigma_{rs} = \nu_{t,u} i^{rt} i^{us} + O(n^{-2})$, $r, s, t, u = 1, \ldots, d$, where $\nu_{r,s} = E_Y\{\ell_r(\omega^*; Y)\ell_s(\omega^*; Y)\}$ and i^{rs} is the (r, s) element of the inverse of the expected information matrix $[i_{rs}] = [-\nu_{rs}]$, with $\nu_{rs} = E_Y\{\ell_{rs}(\omega^*; Y)\}$; ω^* is the pseudo-true parameter value such that $\hat{\omega} = \omega^* + o_p(1)$ (see, for example, White (1994)). Note that $[\sigma_{rs}]$ is the asymptotic covariance matrix for $\hat{\omega}$, under a model which could be misspecified. If the model is correctly specified, that is $g(y) = f(y; \omega_0)$ for $\omega_0 = \omega^*$ in Ω, the well-known information identity $\nu_{r,s} = i_{rs}$ holds and we obtain the usual relation $\sigma_{rs} = i^{rs} + O(n^{-2})$.

The improved information criterion (IIC) is defined as a suitable first-order unbiased estimator for a new target quantity

$$\tilde{\eta}(g, f) = E_Y[E_Z\{\log \tilde{f}(Z; \hat{\omega})\}], \tag{2}$$

which is obtained by substituting in (1) the estimative predictive density $f(z; \hat{\omega})$ with $\tilde{f}(z; \hat{\omega})$. Thus, as proved by Vidoni (2003), the IIC criterion selects the model maximizing

$$\Psi_{IIC}(Y;f) = \log f(Y;\hat{\omega}) - \hat{\nu}_{t,r}\hat{\imath}^{rt} + \frac{1}{2}\,\hat{\nu}_{t,u}\hat{\imath}^{rt}\hat{\imath}^{su}(\hat{\nu}_{r,s} - \hat{\imath}_{rs}),$$

with $\hat{\nu}_{r,s}$ and $\hat{\imath}_{rs}$ suitable estimators for $\nu_{r,s}$ and i_{rs}, respectively. It easy to see that the IIC is a modification of the TIC, which corresponds to $\Psi_{TIC}(Y;f) = \log f(Y;\hat{\omega}) - \hat{\nu}_{t,r}\hat{\imath}^{rt}$. Although, when the model is correct, the IIC and the TIC coincide, and correspond to the AIC, we presume that the IIC will usually present a more accurate discriminating ability than the TIC, since it is based on the improved predictive density.

A preliminary analysis on this conjecture (see also Vidoni, 2003), involves a comparative analysis of the theoretical target quantities (1) and (2) or of the corresponding first order approximations

$$\hat{\eta}(g,f) = E_Y\{\log f(Y;\omega^*)\} - \frac{1}{2}\,\nu_{r,s}i^{rs} + O(n^{-1}),$$

$$\tilde{\eta}(g,f) = E_Y\{\log f(Y;\omega^*)\} - \frac{1}{2}\,\nu_{r,s}i^{rs} + \frac{1}{2}\,\nu_{t,u}i^{rt}i^{su}(\nu_{r,s} - i_{rs}) + O(n^{-1}).$$

We expect that $\tilde{\eta}(g,f)$, which is based on an improved estimator for $g(z)$, presents an additional penalization, with respect to $\hat{\eta}(g,f)$, for misspecified models.

3 Variable selection in logistic regression models

In this section we compare the two theoretical criteria, with respect to the problem of variable selection under logistic regression models. Let $Y_1,\ldots,Y_n$ be mutually independent Bernoulli random variables, with true probability μ_{0i} of being 1. Let us consider the candidate logistic regression model specified by the mean $\mu_i = \exp\{\omega^T x_i\}/[1 + \exp\{\omega^T x_i\}]$, with $x_i = (1, x_{i2}, \ldots, x_{id})^T$ a vector of known covariates and $\omega = (\omega_1, \ldots, \omega_d)^T$ a d-dimensional unknown parameter. In this case,

$$\ell_r(\omega;Y) = \sum_{i=1}^{n}(Y_i - \mu_i)x_{ir}, \quad \ell_{rs}(\omega;Y) = -\sum_{i=1}^{n}\mu_i(1-\mu_i)x_{ir}x_{is},$$

for $r,s = 1,\ldots,d$; $\hat{\omega}$ and ω^* are such that, respectively,

$$\sum_{i=1}^{n}(Y_i - \hat{\mu}_i)x_{ir} = 0, \quad \sum_{i=1}^{n}\{E_Y(Y_i) - \mu_i^*\}x_{ir} = 0,$$

for $r = 1,\ldots,d$. Hereafter, the hat and the asterisk stand for evaluation at $\omega = \hat{\omega}$ and $\omega = \omega^*$, respectively. Indeed, straightforward mathematics leads to $E_Y\{\log f(Y;\omega^*)\} = \sum_{i=1}^{n}\mu_{0i}\log\{\mu_i^*\} + \sum_{i=1}^{n}(1-\mu_{0i})\log\{1-\mu_i^*\}$, $i_{rs} = \sum_{i=1}^{n}\mu_i^*(1-\mu_i^*)x_{ir}x_{is}$ and $\nu_{r,s} = \sum_{i=1}^{n}\mu_{0i}(1-\mu_{0i})x_{ir}x_{is}$. Note that a sufficient condition assuring $\Psi_{IIC}(Y;f) = \Psi_{TIC}(Y;f) = \Psi_{AIC}(Y;f)$ is that the model is "true" or it is a suitable generalization of the true one.
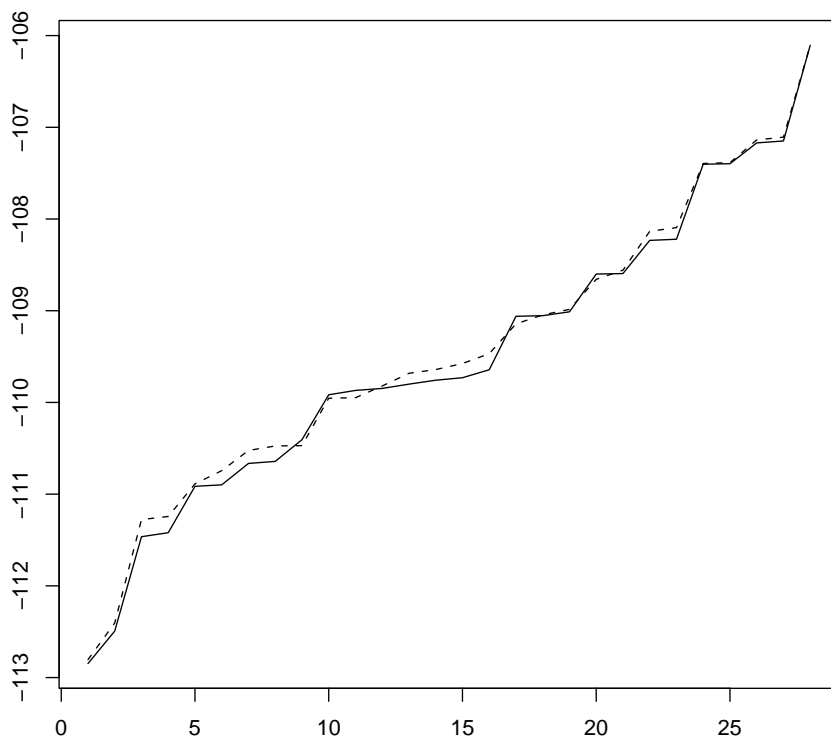
FIGURE 1. Theoretical criteria $\hat{\eta}(g, f)$ (dashed line) and $\tilde{\eta}(g, f)$ (solid line) for the 28 alternative logistic regression models with $d = 7$ parameters.

Let us assume that the true model is a logistic regression with d_0 covariates, chosen in a set of potential covariates, and that all the candidate logistic regression models have $d = d_0$ covariates, which may differ from those specifying the true one. In this situation, since all the alternative models have the same number of unknown parameters, the penalization given by the AIC is fixed to d_0. Thus, model selection using the AIC involves in fact only the maximized log-likelihood $\log f(Y; \hat{\omega})$. Here, we aim to compare the discriminating ability associated to the two alternative theoretical target quantities $\hat{\eta}(g, f)$ and $\tilde{\eta}(g, f)$.

We shall consider the birthweight data, provided by Hosmer and Lemeshow (2000). The response variable is the indicator of birth weight less than 2.5 kg, there are 8 explanatory variables and the number of observations is $n = 189$. Let us assume that the true logistic regression model has $d_0 = 7$ parameters, namely, the intercept and those ones related to the covariates named "lwt", "race", "smoke", "ptl", "ht", "ui".

The true parameter values are set equal to maximum likelihood estimates

obtained by the original data. We define 28 alternative logistic regression models with $d = 7$ parameters, with the intercept included. Figure 1 plots the values of the (approximated) theoretical criteria $\hat{\eta}(g, f)$ and $\tilde{\eta}(g, f)$, in ascending order, for the 28 alternative regression models. As expected, the two criteria select the true model (here the 28th) and $\tilde{\eta}(g, f)$ usually presents an additional penalization, with respect to $\hat{\eta}(g, f)$, for the misspecified models. Thus, the model selection criterion based on the improved predictive density has, in this case, a better discriminating ability. Similar results may be obtained if we consider different true models. An extended analysis comparing the two alternative criteria is given by Vidoni (2003).

References

Akaike, H. (1973). Information theory and extension of the maximum likelihood principle. In: *Second Symposium on Information Theory*, N.B. Petron and F. Caski (Eds.). 267-281, Budapest: Akademiai Kiado.

Corcuera, J.M. and Giummolè, F. (2000). First order optimal predictive densities. In: *Applications of Differential Geometry to Econometrics*, P. Marriott and M. Salmon (Eds.). 214-229, Cambridge: Cambridge University Press.

Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression* (2nd Ed.). New York: Wiley

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, **57**, 120-125.

Shibata, R. (1989). Statistical aspects of model selection. In: *From Data to Model*, J.C. Willems (Ed.). 215-240, New York: Springer-Verlag.

Vidoni, P. (2003). Improved predictive model selection. *Technical Report 8.03*, Department of Statistics, University of Udine.

White, H. (1994). *Estimation, Inference and Specification Analysis*. New York: Cambridge University Press.