Accelerated Failure Time Model for Arbitrarily Censored Data With Smoothed Error Distribution

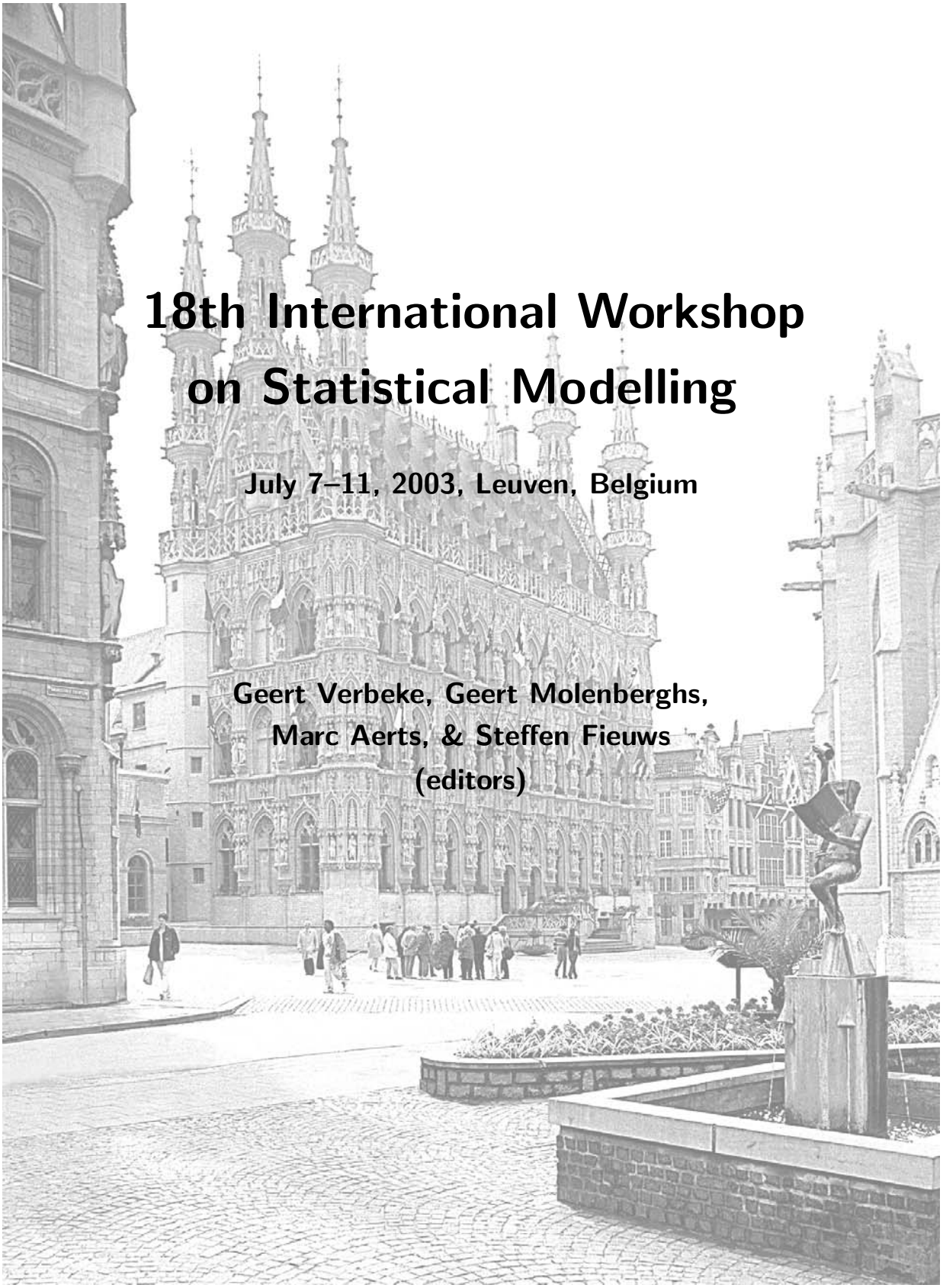
Article *in* Journal of Computational and Graphical Statistics · September 2005

DOI: 10.1198/106186005X63734

3 authors, including:

Arnošt Komárek
Charles University in Prague
10 PUBLICATIONS **236** CITATIONS
SEE PROFILE

Emmanuel Lesaffre
Erasmus MC
475 PUBLICATIONS **15,730** CITATIONS
SEE PROFILE



18th International Workshop on Statistical Modelling

July 7–11, 2003, Leuven, Belgium

Geert Verbeke, Geert Molenberghs,
Marc Aerts, & Steffen Fieuws
(editors)

18th International Workshop on Statistical Modelling

July 7–11, 2003, Leuven, Belgium

Geert Verbeke, Geert Molenberghs,
Marc Aerts, & Steffen Fieuws
(editors)

ISBN 90-807934-1-8

The Editors

Geert Verbeke
Biostatistical Centre
Catholic University of Leuven
Kapucijnenvoer 35
B-3000 Leuven, Belgium

Geert Molenberghs
Center for Statistics
Limburgs Universitair Centrum
Universitaire Campus, Building D
B-3590 Diepenbeek, Belgium

Marc Aerts
Center for Statistics
Limburgs Universitair Centrum
Universitaire Campus, Building D
B-3590 Diepenbeek, Belgium

Steffen Fieuws
Biostatistical Centre
Catholic University of Leuven
Kapucijnenvoer 35
B-3000 Leuven, Belgium

The correct bibliographic citation for this proceedings is as follows:

Author(s) (2003). Title. 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. 000–000.

Scientific Committee

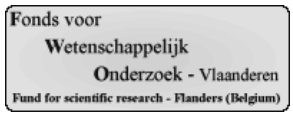
- M.Aerts (Diepenbeek, Belgium)
- A.Biggeri (Florence, Italy)
- L.J.Brant (Baltimore, The United States)
- T.Burzykowski (Diepenbeek, Belgium)
- P.Delaportas (Athenes, Greece)
- P.Diggle (Lancaster, The United Kingdom)
- M.Hubert (Leuven, Belgium)
- E.Lesaffre (Leuven, Belgium)
- G.Molenberghs (Diepenbeek, Belgium)
- V.Nunez-Anton (Bilbao, Spain)
- J.Palmgren (Stockholm, Sweden)
- M.Stasinopoulos (London, The United Kingdom)
- L.Ugarte (Pamplona, Spain)
- M.Vandebroek (Leuven, Belgium)
- G.Verbeke (Leuven, Belgium)

Local Organizing Committee

- M.Aerts (Diepenbeek, Belgium)
- K.Bogaerts (Leuven, Belgium)
- L.Bruckers (Diepenbeek, Belgium)
- A.Carbonez (Leuven, Belgium)
- S.Fieuws (Leuven, Belgium)
- A.Ghesquiere (Leuven, Belgium)
- E.Lesaffre (Leuven, Belgium)
- M.Machiels (Diepenbeek, Belgium)
- G.Molenberghs (Diepenbeek, Belgium)
- J.Rongy (Leuven, Belgium)
- G.Verbeke (Leuven, Belgium)

Sponsors

- Arnold, The United Kingdom
- Cosinus Computing BV
- Fund for Scientific Research – Flanders, Belgium
- Janssen-Cilag, Belgium
- Johnson & Johnson, Division of Janssen Pharmaceutica, Belgium
- John Wiley & Sons, The United Kingdom
- Merck & Co., Belgium
- Novartis, Belgium
- SAS Institute, Belgium
- Solvay



Preface

This proceedings volume contains the papers presented at *The 18th International Workshop on Statistical Modelling* held in Leuven, Belgium, July 7–11, 2003. The workshop aims to bring together researchers and all those interested in the development of statistical models and in their applications in the widest sense. It arose out of the idea of having a forum for presenting and discussing advances in statistical modelling and stimulating international collaborative work. The main focus is the annual meeting (usually held in July) where a wide range of non-theoretical papers from a wide range of areas in addition to considering theoretical contributions are covered.

The International Workshop on Statistical Modelling has been held in Europe and the USA for the past 18 years. The workshop arose out of two GLIM conferences in the U.K. in London (1982) and Lancaster (1985), and from a number of short courses organised by Murray Aitkin and held at Lancaster in the early 1980s, which attracted many European statisticians interested in Generalised Linear Modelling. At this time, a group of Austrian, Italian and British statisticians saw both the opportunity and the need for a regular meeting of Europeans that would focus on various aspects of statistical modelling in an informal workshop 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 both motivated by real life data and which also 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. Special attention is given to student contributions, and an award for the best student presentation is given. The scientific programme is characterised by having invited lectures and a pre-workshop short course, contributed papers, posters and software demonstrations.

Since the first meeting in Innsbruck in 1986, the workshop has grown substantially, and now regularly attracts over 200 participants. There has been a strong effort to bring each new meeting to a different European country. The scope of the workshop is now much broader, reflecting the growth in the subject of statistical modelling over ten years. The number of submitted papers has grown with the number of participants, but parallel sessions have been avoided, allowing everyone both to learn and to contribute. Poster sessions are now held, and software demonstrations and displays are organised. One change is that the workshops have become more international in nature. Participants now attend from all corners of the globe, and workshops have travelled around Europe: Innsbruck (1986), Perugia (1987), Vienna (1988), Trento (1989), Toulouse (1990), Utrecht (1991), Munich (1992), Leuven (1993), Exeter (1994), Innsbruck (1995), Orvieto (1996), and Biel/Bienne (1997) - to the USA - New Orleans (1998) - and back to Europe - Graz (1999), Bilbao (2000), Odense (2001), Chania (2002). Future workshops will be organized in Florence (2004) and Australia (location to be specified, 2005).

After 10 years, the workshop is back in Leuven, as a joint organization of the Biostatistical Centre of the K.U.Leuven and the Center for Statistics of the Limburgs Universitair Centrum. The scientific programme consists of invited papers, oral contributions as well as poster contributions. We very much appreciate the efforts of the scientific committee in the selection of the invited speakers and the oral contributions. We thank the invited speakers, Ron Brookmeyer (The Johns Hopkins University, U.S.A.), Chris Chatfield (The University of Bath, U.K.), Marie Davidian (North Carolina State University, U.S.A.), Anastasios Tsiatis (North Carolina State University, U.S.A.), and Henry Wynn (London School of Economics, U.K.) for accepting the invitation to present a one hour state of the art lecture in their specific fields of expertise. Abstracts of these presentations are included in the first part of this volume. Further, we very much appreciate the efforts of Brian Marx (Louisiana State University, U.S.A.) and Paul Eilers (University of Leiden, The Netherlands) for their one-day short course entitled 'Smoothing for Smarties.' Finally, our special thanks are dedicated to all authors who contributed to the second and main part of this proceedings volume, for participating in the workshop, and for carefully preparing their manuscripts. Finally, we wish all participants a pleasant stay in the historic city of Leuven, and a very fruitful scientific meeting.

Geert Verbeke
Geert Molenberghs
Marc Aerts
Steffen Fieuws

Leuven, May 2003

Contents

Part I: Abstracts of Invited Papers

Part II: Presented Papers

Part I

Abstracts of Invited Papers

Statistical Models for Anthrax Outbreaks

Ron Brookmeyer

[1] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
Baltimore, Maryland USA

In the fall of 2001 an outbreak of inhalational anthrax occurred in the United States that was the result of bioterrorism. Letters contaminated with anthrax spores were sent through the postal system. In response to the outbreak, public health officials treated over 10,000 persons with antibiotic prophylaxis in the hopes of preventing further morbidity and mortality. No persons receiving the antibiotics subsequently developed disease. The question arises how many cases of disease may actually have been prevented by the public health intervention of antibiotic prophylaxis. In this paper, a statistical model is developed to answer this question by relating the dates of disease onset, initiation of antibiotic prophylaxis, and exposure to the anthrax spores, to the incubation period distribution. An important complication is that the date of exposure to the anthrax spores was unknown for a cluster of cases in Florida because the contaminated letter was never found.

A general likelihood function for a multi-common source outbreak is developed where the dates of exposure to the source (e.g. anthrax spores) may or may not be known. Estimates of the incubation period distribution are derived from an outbreak in Sverdlovsk, Russia. The results are applied to the 2001 U.S. outbreak to estimate jointly the date the Florida cases were exposed to the contaminated letter, and the numbers of cases of disease that may have been prevented in the three main clusters in New Jersey, Florida and Washington, D.C. The model is extended to allow a phase-in time period during which antibiotics are distributed. The sensitivity of the estimates to the assumed incubation period is investigated. Properties of the estimators particularly when the outbreak sizes are small are evaluated by simulation. We find that antibiotics may have cut in half the number of cases of disease. Sensitivity analyses indicate that even in the absence of antibiotic prophylaxis the outbreak would not likely have been more than 50 cases. The results underscore the importance of early detection of outbreaks together with targeted and effective public health control measures.

Keywords: Anthrax; Epidemiology; Infectious disease; Likelihood.

Model Selection, Data Mining and Model Uncertainty

Chris Chatfield[1]

[1] Department of Mathematical Sciences, University of Bath, Bath, UK, BA2 7AY

Different methods for selecting an appropriate model are briefly reviewed. Data mining, or data dredging, arises when large numbers of models are tried on the same data. The effects of this, such as model-selection bias, are still not widely understood and some remarks are made on model uncertainty.

Keywords: Akaike's information criterion; Bayesian information criterion; Data dredging; Principle of parsimony.

An extended version can be found on page 79.

"Semiparametric" Approaches for Inference in Joint Models for Longitudinal and Time-to-Event Data

Marie Davidian[1]

[1] Department of Statistics, North Carolina State University, Raleigh, USA

A common objective in longitudinal studies is to characterize the relationship between a longitudinal response process and a time-to-event. Considerable recent interest has focused on so-called joint models, where models for the event-time distribution (typically proportional hazards) and longitudinal data are taken to depend on a common set of latent random effects, which are usually assumed to follow a multivariate normal distribution. A natural concern is sensitivity to violation of this assumption. We review the rationale for and development of joint models and discuss two modeling and inference approaches that require no or only mild assumptions on the random effects distribution. In this sense, the models and methods are semiparametric. The methods will be demonstrated by application to data from an HIV clinical trial.

Keywords: Longitudinal data; Time-to-event data; Semiparametric.

Efficient Estimation of The Mean of A Time-Lagged Variable Subject to Right Censoring

Anastasios A. Tsiatis

[1] Department of Statistics, North Carolina State University, Raleigh, USA

In many clinical trials, the endpoint of interest may not be available immediately, but rather evolves over time. Examples are numerous. Survival time is clearly such an example, but also cost-of-care, quality-adjusted lifetime, or even dichotomous response such as whether viral load falls below detectable limits after treatment for AIDS patients are also examples of time- lagged responses. The lag time may be part of the biological process or due to administrative delays. Because patient entry is staggered and follow-up is of limited duration, some of the response variables will be missing due to censoring of the lag time. We will show how the theory of inverse probability weighting of complete cases developed by Robbins and Rotnitzky can be used to derive consistent estimators for the mean of a time-lagged variable. We will also show how to use additional information collected during the study to increase effciency.

Keywords: Efficient estimation; Time-lagged variable; Right censoring.

Computational Algebraic Methods for Discrete Statistical Models

Henry P. Wynn

[1] London School of Economics, UK.

Algebraic statistics is the name given to the use of computational algebraic methods in statistics covering in particular graphical models and various independence and conditional independence structures. The use of algebra comes from various sources. First by interpolating the probability mass function or its logarithm over the support using Gröber basis methods one obtains unique forms of polynomial or exponential models. This immediately copes with difficult support problems such as structural zeros in contingency tables.
Second, complex factorisations can be expressed algebraically. Thus for conditional independence of X_1 and X_2 on X_3, in the binary case, we have the exponential form:

$$p(x, y, z) = \exp(\phi_{000} + \phi_{100}x_1 + \phi_{010}x_2 + \phi_{001}x_3 + \phi_{101}x_1x_3 + \phi_{011}x_2x_3)$$

By setting

$$t_\alpha = \exp(\phi_\alpha)$$

for each multi-index α we obtain another algebraic formulation. Then eliminating the t_α we obtain the "toric ideal" representation.
The representation comes from the special choice of multi-indices α defining the original factorisation. There is a very close connection between the set of multi-indices and certain inclusion-exclusion identities based on the sets. For the above conditional independence, for example we have:

$$\{123\} = \{13\} + \{23\} - \{3\}$$

The key to the use of such identities in the modelling environment is certain projection operators based on conditional expectations.
In summary, many factorisations such as graphical models, junction trees and similar structures can be classified using such identities. Being able to move between the different algebraic formulations of models and submodels is revealing. The work summarises collaboration with co-workers

especially: G Pistone (Torino), E Riccomagno (Warwick).

Keywords: Algebraic Methods, Discrete Statistical Models.

Part II

Presented Papers

Two Lack of Fit Tests for Multiple Logistic Regression

M. Aerts[1], G. Claeskens[2], J. Hart[2], E. Moons[3], and G. Wets[3]

[1] Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium
[2] Department of Statistics, Texas A & M University, College Station, Texas 77843, U.S.A.
[3] Data Analysis and Modeling Group, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium

Abstract: Several methods have been developed to asses the fit of a regression model. Many lack of fit tests however focus on the simple regression setting. Here we propose two tests which are completely different in nature, but which both are promising especially in the case of a multiple regression model with several potential explanatory variables.

Keywords: Bayes information criterion; Classification trees; Lack of fit; Posterior distribution; Recursive partitioning.

1 Introduction

There is a variety of techniques and methods available for testing lack of fit in regression models, see e.g. Hart (1997). Here we focus on the special case of multiple logistic regression. Other related work covering this setting includes Brown (1982), le Cessie and Van Houwelingen (1991, 1993, 1995), Aerts, Claeskens and Hart (2000).
Consider a binary response Y on N subjects and a logistic regression model $\text{logit}\{P(Y = 1)\} = g(\boldsymbol{x})$ with g some unknown regression function and $\boldsymbol{x} = (x_1, ..., x_p)$ the covariate vector (possibly containing mixed continuous and categorical variables). The null hypothesis states

$$H_0 : g(\boldsymbol{x}) = g(\boldsymbol{x}; \boldsymbol{\theta}) \tag{1}$$

as the correct model for our data, where $g(\boldsymbol{x}; \boldsymbol{\theta})$ is a specified parametric regression function and $\boldsymbol{\theta}$ an unknown parameter vector. In case all explanatory variables are categorical and if sparseness is no issue, we can rely on the Pearson goodness-of-fit test, given by

$$\chi^2 = \sum_{i=1}^{I} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1 - \hat{\pi}_i)}, \tag{2}$$

where n_i is the number of subjects in covariate pattern i, y_i the number of observed events and $\hat{\pi}_i$ the fitted probability based on the null model M_0 in the i-th of all I possible combinations. In case there are one or more continuous explanatory variables, the asymptotic distribution of Pearson's chi-squared and the deviance test is not applicable anymore. The following two tests offer a solution in this setting. A first approach is a modification of the well-known Hosmer-Lemeshow test. The second method is a Bayesian-motivated test and can carried out in either Bayesian or frequentist fashion.

2 A Tree Based Test

The Hosmer-Lemeshow test has the same form as the Pearson test statistic but the grouping is different and typically based on the so-called deciles of risk. More precisely, the first group contains those subjects (10 % of the sample size) with the smallest estimated (under the null hypothesis) probabilities, etc. Since this grouping is based on the fitted null model, this approach is expected to have nonoptimal power characteristics.
The grouping proposed here is based on a flexible nonparametric model, the classification tree. The test statistic actually measures the discrepancy between the parametric null model and the classification tree as its unrestricted nonparametric counterpart. In general, a tree consists of different layers of nodes (implying a grouping, splitting of the sample). It starts from the root node in the first layer, containing all data. Using an impurity criterion (maximizing the homogeneity), this parent node is split into two daughter nodes on the second layer. This partitioning process continues until a stopping criterion is reached. The tree is then pruned to an optimal sized tree during the pruning process. It is the grouping of this final tree which is used to define a Hosmer-Lemeshow like test statistic.
As a consequence of the data-driven grouping procedure, the final groups might be highly unbalanced with some of the groups containing only a few observations. To improve the distributional behaviour of the test statistic (2), the tree test can be based on the Cressie and Read (1984) family of power divergence statistics, i.e.

$$T_{CR} = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{I} \left\{ y_i \left(\left(\frac{y_i}{n_i \bar{\pi}_i} \right)^{\lambda} - 1 \right) + (n_i - y_i) \left(\left(\frac{n_i - y_i}{n_i (1 - \bar{\pi}_i)} \right)^{\lambda} - 1 \right) \right\} \quad (3)$$

with I the number of final nodes resulting from the partitioning and pruning process, $\bar{\pi}_i$ the proportion of observed events in the ith group and $-\infty < \lambda < \infty$. For $\lambda = 1$ it equals the Pearson based formulation (2). Cressie and Read (1984) recommend the statistic with $\lambda = \frac{2}{3}$, which they found less susceptible to effects of sparseness.
Because the cells in the contingency table are random, the distribution of the tree based test cannot be obtained from a straightforward application

of the usual theory for chi-squared goodness of fit tests (Moore and Spruill 1975). Simulations show that a chi-squared distribution with $2 \times g - p$ degrees of freedom is a reasonable choice. Next to this approximate distribution, one can always simulate a null distribution by a parametric bootstrap method. For a detailed discussion, see Moons, Aerts and Wets (2002).

3 A Bayesian Motivated Test

A version of this test was proposed by Hart (1997). The idea is very simple. Consider a sequence of models for $g(\boldsymbol{x})$ of varying dimensions, one of which is the parametric null model $g(\boldsymbol{x};\boldsymbol{\theta})$. The posterior probability, $\tilde{\pi}_n$, of the null model is computed, and if this probability is sufficiently low, the null model is rejected. A sequence of constants a_n can be determined such that $a_n(1-\tilde{\pi}_n)$ converges in distribution to a nondegenerate random variable when H_0 is true and the sample size n tends to ∞. This allows the frequentist to conduct a valid large sample test of given size based on $a_n(1-\tilde{\pi}_n)$. There are several choices for the sequence of alternative models M_j. Here, we consider nested models ($M_j \subset M_{j+1}$) and singleton models that contain only one more parameter than the null model M_0.
Applying Schwartz 's (1978) approximation, we get the following approximation of the posterior probability

$$P(M_0|\boldsymbol{y}) \approx \frac{1}{1+\sum_{j=1}^{K}\exp(BIC_j - BIC_0)} \stackrel{\text{def}}{=} \pi_{BIC}.$$

where $BIC_j = \log L_j - m_j \log n/2$, the Bayes Information Criterion of model M_j (L_j the likelihood function at the MLE and m_j the dimension of model M_j).
Under certain regularity conditions and for a finite number K of alternative models, , it can be shown that $n^{\frac{1}{2}}(1-\pi_{BIC}) \rightarrow_{\mathcal{D}} \exp(V_1/2)$ for the nested models and $n^{\frac{1}{2}}(1-\pi_{BIC}) \rightarrow_{\mathcal{D}} \sum_{k=1}^{K}\exp(V_k/2)$ for the singletons, where $V_1,\ldots,V_K$ are independent χ^2_1 random variables and K is the total number of alternative singleton models. For more details on the limiting null distribution (including finite sample corrections and the case in which the number of alternative models tends to ∞ with n) and on the power against local alternatives, see Aerts, Claeskens and Hart (2003).

4 Data Example and Discussion

The data set used in this analysis comes from the Project on Preterm and Small-for-Gestational-Age Infants in the Netherlands (POPS), a Dutch follow-up study on preterm infants by Verloove and Verwey (1988), see also le Cessie and van Houwelingen (1991) . Data were collected on 1338 infants, born in 1983 in The Netherlands with a gestational age of less than 32

TABLE 1. *Test results POPS data: p-values for three null models.*

Test	x_1, x_1^2, x_2	x_1, x_2, x_2^2	x_1, x_1^2, x_2, x_2^2
B_S	0.000	0.000	0.126
B_N	0.006	0.000	0.138
T_{CR}	0.012	0.044	0.090
HL	0.125	0.002	0.207
CVH	0.02	-	0.45
BR	0.01	-	0.06
$ACH1$	-	-	0.07
$ACH2$	-	-	0.02

completed weeks and/or a birthweight of less than 1500 g. After deleting the observations with missing data, a data set of 1310 infants remained. We consider the situation after 2 years. The response variable Y indicates whether or not the infant has died within 2 years or has survived but with a major handicap. The explanatory variables are gestational age (X_1) and weight of the babies at birth (X_2). As an illustration, we consider each of the following models as null model: model 1 with x_1, x_1^2, x_2, model 2 with x_1, x_2, x_2^2, and model 3 with x_1, x_1^2, x_2, x_2^2. Table 1 shows the results for the tree-based and the Bayesian motivated test and compares them with the results from several other tests from literature. The first four lines shows p-values for the singleton and nested Bayesian motivated test (B_S and B_N respectively) using a sequence of alternative models including up to fifth order main and interaction effects, the tree-based test based on the Cressie-Read statistic (T_{CR}) with pruning up to 15 terminal nodes, and the Hosmer-Lemeshow test (HL) based on deciles of risk. All p values were simulated using the parametric bootstrap (1000 runs).
The last four lines show some analogous results from other test statistics proposed in literature: a kernel based goodness of fit method (CVH) proposed by le Cessie and Van Houwelingen (1991, 1993), the Brown statistic (BR, Brown 1982, see also le Cessie and Van Houwelingen 1993) and an order selection score test ($ACH1$) and the value of a score based AIC criterion ($ACH2$) as reported by Aerts, Claeskens and Hart (2000).
The p-values in Table 1 show that there is clear evidence against any model without both quadratic terms (model 1 and 2). Only the HL test does not reject model 1. As also discussed in Aerts, Claeskens and Hart (2000), there is some evidence against model 3 with both quadratic terms, but the different test results disagree. The HL test seems to have less power than the tree based test T_{CR}, which has been confirmed by simulations (see Moons, Aerts and Wets 2002). On the other hand this latter test has less convincing results for the simpler null models. Especially the Bayes motivated tests reject model 1 and 2 very strongly. Of course, such conclusions

are premature. Extensive simulations are needed to shed more light on the power characteristics of the different test statistics. An appealing property of the Bayes motivated test is that it can be easily implemented for more complex likelihood models (like e.g. for clustered data). The tree-based test is promising in settings with huge datasets (like in data mining).

References

Aerts, M., Claeskens, G., and Hart, J.D. (2000). Testing lack of fit in multiple regression. *Biometrika*, **87**, 405–424.

Aerts, M., Claeskens, G., and Hart, J.D. (2003). Bayesian-motivated tests of function fit and their asymptotic frequentist properties. Submitted.

Brown, C.C. (1982). On a goodness of fit test for the logistic model based on score statistics. *Communications in Statistics - Theory and Methods*, **11**, 1087–1105.

Cressie, N. and Read, T. (1984). Multinomial goodness of fit tests. *Journal of the Royal Statistical Society, Series B*, **46**, 440–464.

Hart, J.D. (1997). *Nonparametric Smoothing and Lack-of-fit Tests*. New York: Springer-Verlag.

le Cessie, S. and Van Houwelingen, C. (1991). A goodness of fit test for binary regression models, based on smoothing methods. *Biometrics*, **47**, 1267–1282.

le Cessie, S. and Van Houwelingen, C. (1993). Building logistic models by means of a non parametric goodness of fit test: a case study. *Statistica Neerlandica*, **47**, 97–109.

le Cessie, S. and Van Houwelingen, C. (1995). Testing the fit of a regression model via score tests in random effects models. *Biometrics*, **51**, 600–614.

Moons, E., Aerts, M., and Wets, G. (2002). Tree based lack-of-fit tests. Submitted.

Moore, D.S. and Spruill, M.C. (1975). Unified large-sample theory of general chi-squared statistics for tests of fit. *Annals in Statistics*, **3**, 599–616.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals in Statistics*, **6**, 461–464.

Verloove, S. and Verwey R. Y. (1988). Project on preterm and small-for-gestational age infants in the Netherlands, 1983, University Microfilms International, no. 8807276. Ann Arbor, MI.

Evolving Classification and Regression Trees

Claudio Agostinelli and Irene Poli

[1] Department of Statistics, University Ca' Foscari, Campiello S. Agostin, S.Polo 2347, 30125 Venice, Italy, claudio@unive.it, irenpoli@unive.it

Abstract: In this paper we introduce a new procedure to build the Classification and Regression Trees. The procedure called ECART is based on a genetic algorithm.

Keywords: Classification; Genetic algorithms; Regression trees.

1 Introduction

Classification And Regression Trees (CART) (Breiman et al, 1984) is a popular procedure for regression and classification problems where high dimensionality and a non–linear optimization criterion are involved. In particular, CART is a nonparametric statistical method developed to build models with a tree–based structure.
Considering a regression problem, where $\mathbf{X} = (X_1, \cdots, X_p)$ is the input space and Y is the response variable, the CART algorithm adopts a binary recursive partitioning startegy: the input space R_0, with $\mathbf{X} \in R_0$, is divided into two regions R_1 and R_2 by a split (i, a) on the variable X_i at the split point a. The procedure selects i and a so that replacing the parent region R_0 with the two regions R_1 and R_2 yields minimal empirical risk. The algorithm proceeds recursively on the daughter regions until a very large number of regions are achieved. Model selection criteria are then applied for stopping the process.
In this procedure to model the relation between Y and X we are asked to choose: i) which and how many variables to introduce in the model, ii) in which order the variables should appear, iii) the number and the position of the split points, and iv) the associated regression parameters.
Choosing all these elements of the model to minimize empirical risk is a hard combinatorial problem, and CART represents an approximate solution based on a recursive partitioning approach.
CART is a simple algorithm to implement and fast to compute. Unfortunately CART also produces sub–optimal solutions and can be an unstable procedure (removing or adding a few observations may change the tree structure). Recently, Breiman (2001) introduced the Random Forest (RF) to avoid instability.

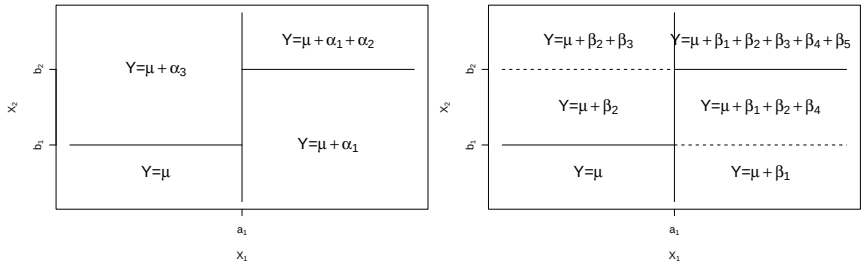


FIGURE 1. *The partition generated by the tree model [1] (left) and by the model [2] (right).*

In this work we adopt an evolutionary approach to build the CART method. We design a Genetic algorithm (GA) which evolves the partition of the input space, and then achieves how many and which split points for each variable considered. We introduce a two stage genetic algorithm where the first stage is designed to evolve the number of the split points in each X variable and the second stage the position of the split points.
To introduce this algorithm let us consider I_{i,a_j} an indicator function such that

$$I_{i,a_j} = \begin{cases} 1 & X_i > a_j \\ 0 & X_i \leq a_j \end{cases} \qquad 1 \leq i \leq p \text{ and } j \geq 0 \ ,$$

and $\bar{I}_{i,a_j} = (1 - I_{i,a_j})$ the complement of I_{i,a_j}.
The simple regression model (with the tree representation as in Figure 1):

$$Y = \mu + \alpha_1 I_{1,a_1} + \alpha_2 I_{1,a_1}\ I_{2,b_2} + \alpha_3 \bar{I}_{1,a_1} I_{2,b_1} + \varepsilon \tag{1}$$

with $\mu, \alpha_1, \alpha_2, \alpha_3, a_1, b_1, b_2$ as unknown constants and ε as a normal random variable with mean zero and variance σ^2, is reformulated in the following way

$$Y = \mu + \beta_1 I_{1,a_1} + \beta_2 I_{2,b_1} + \beta_3 I_{2,b_2} + \beta_4 I_{1,a_1} I_{2,b_1} + \beta_5 I_{1,a_1} I_{2,b_2} + \varepsilon' \tag{2}$$

This reformulated model [2] is related to the regression tree model [1]: a few constraints are introduced on the β's (see also Figure 1), such as $\beta_1 = \alpha_1$, $\beta_2 = \alpha_3$, $\beta_3 = 0$, $\beta_4 = -\beta_2 = -\alpha_3$, $\beta_5 = \alpha_2$.
Models with this formulation present the following advantages:

1. the variables and their split points can be introduced or cancelled from the model without modified the remaining structure;
2. the order of the variables is not relevant and this simplifies and speeds up the search of a sub–optimal model;
3. the cardinality of the set of models grows very slowly.

The formulation (2) of the regression model allow us to build a genetic algorithm which evolves the possible candidate solutions to our problem, chosen in a very large sets of possible solutions.

2 The Genetic Algorithm

Genetic algorithms (Holland, 1975, and Goldberg, 1989) are powerful and flexible tools for search and optimization problems. They are based on the mechanics of natural selection and they are particularly suitable for optimization problems involving discrete parameters. They have been successfully applied in a large variety of fields and problems, including the selection of statistical models (Minerva and Poli, 2001).
In this work we design a two-stage genetic algorithm in order to choose the number of split points and their values. In fact, the first stage creates and evolves a population of candidate numbers to be considered as the number of split points of each variable of the input set. The second stage creates and then evolves a population of possible split points for each variable of the input set. The algorithm works with a transformed form of the Akaike criterion (AIC) as a fitness function to select the optimal models.
The implementation of the genetic algorithm follows these steps:

1. Select random values from a discrete Uniform distribution with support $0, 1, \cdots, N_i$, where N_i represents the maximum number of split points for the i-th input variable with $i = 1, \cdots, p$, and create a population of individuals $\mathbf{n_s}$, $s = 1, \cdots, S$ where S is the size of the population; each individual defines a possible set of the number of the split points of the input space variables, that is $\mathbf{n}_s = (n_{s1}, n_{s2}, \cdots, n_{sp})$; encode each individual with the reflected Gray code;

2. For each individual $\mathbf{n}_s$ create a random population of new individuals $\mathbf{v}_k$, $k = 1, \cdots, K$, where K is the size of the population. Each individual is a p vector whose elements are vectors of variable size, that is $\mathbf{v}_k = (\mathbf{v}_{k1}, \mathbf{v}_{k2}, \cdots, \mathbf{v}_{kp})$, with $\mathbf{v}_{ki} = (v_{ki,1}, v_{ki,2}, \cdots, v_{ki,n_{si}})$; each element of the vector $\mathbf{v}_{ki}$, represents a rank value identifying the corresponding value assumed by the i-th input variable; encode each individual with the reflected Gray code;

3. Compute the fitness function values;

4. Select the values of the parameters for the selection, crossover and mutation operators and perform the genetic operators as defined above;

5. Set g=g+1; if $g \leq N_g$ then go to step 3;

6. Assign to each individual n_s the best fitness function achieved with the last generation of the previous GA (which identifies the best choice of the values of the input variables given the number of the splits: steps 2–5;

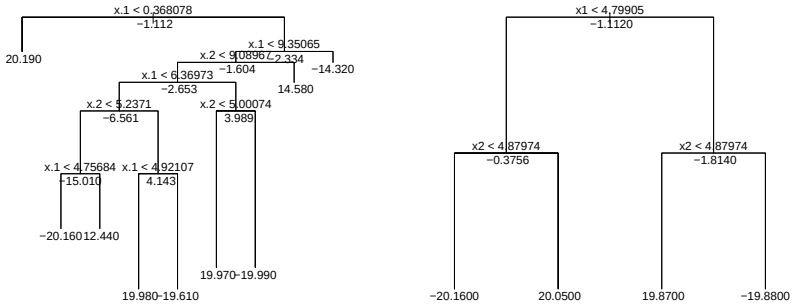


FIGURE 2. *The tree generated with CART (left) and ECART (right).*

7. Select the values of the parameters for the selection, crossover and mutation operators and perform the genetic operators as defined above;
8. Set G=G+1; if $G \leq N_G$ then go to step 2;
9. Return the best $\mathbf{n}_s$ and the associated $\mathbf{v}_k$ from the last generation.

This algorithm has been applied to few set of data giving very encouraging results.

3 An illustrative Example

We present the results achieved with a sample of observations from data with the following model:

$$Y = -20 + 40\ I_{1,5} + 40\ I_{2,5} - 80\ I_{1,5}I_{2,5} + \varepsilon$$

and $X_1, X_2 \sim \mathrm{U}(0, 10)$, $\varepsilon \sim \mathrm{N}(0; 1)$ and a sample of size 129 from it.
The tree structure from CART is reported in Figure 2 (left). The suggested tree has several branches that we can not prune without losing the correct splits. From the tree it is really difficult to recover the true structure of the data.
We run the genetic algorithm described above using $S = K = 20$. The parameter of crossover is set to 0.95 and that of mutation is set to 0.02. We run $N_g = N_G = 10$ generations. The number of possible partitions we consider is of order 10^{28} and with the algorithm, at the end, we explore 40000 of them, a very small number. The best solution is found after exploring about 12000 partitions (5-th generation).
Table 1 compares CART with our procedure (ECART). We compute the deviance ($Dev(R)$) of the residuals and deviance ($Dev(Y)$) of the dependent variable.

TABLE 1. *Results for CART and ECART.*

	True model	CART	ECART
Dev(R)	125.113	4180.693	125.113
Fitness ($\times 10^{-3}$)	11.529	2.745	11.529
Misclassified (%)	0.000	2.326	0.000
Dev(Y)		51493.370	

References

Agostinelli, C. and Poli, I. (2003). Evolutionary computation for classification and regression trees. *Working Paper* 2003.3, Department of Statistics, University of Venice.

Breiman, L. (2001). Random forests. *Machine Learning*, **1**, 5–32.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees*. Chapman & Hall/CRC.

Goldberg, D.E. (1989). *Genetic Algorithms in Search, Optimization & Machine Learning*. Addison-Wesley Publishing Company.

Holland, J.H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press.

Minerva, T. and Poli, I. (2001). Building arma models with genetic algorithms. In: *Applications of Evolutionary Computing*, E.J.W. Boers et al, editor. Berlin: Springer-Verlag.

A non-PH Accelerated Hazard Model for Analyzing Interval Censored Trial Data

Yasin Al-Tawarah[1] and Gilbert MacKenzie[1]

[1] Centre for Medical Statistics, Keele University, Keele, Staffordshire ST5 5BG, UK. Email: g.mackenzie@keele.ac.uk

Abstract: In longitudinal studies with a set of continuous or ordinal repeated response variables it may be convenient to summaries the outcome as a threshold event. Then, the time to this event becomes of interest. In this paper we obtain the general likelihood for the unknown parameters when the underlying survival model is parametric and the survival times are interval-censored. We investigate the use of a member of the Generalized Time Dependent Logistic family of survival distributions (MacKenzie, 1996) which is a non-PH Accelerated Hazard Model and has a logistic baseline hazard function. We use simulation to investigate how inference on the treatment parameter is compromised by using the mis-specified likelihood, which treats the interval-censored survival times as if they were exact

Keywords: Interval censoring; Logistic survival; Non-PH model; Accelerated hazard; Mis-specified likelihood.

1 Introduction

In classical survival analysis, the exact time to event is usually known. However, in longitudinal clinical trials where outcome is a continuous or ordinal variable measured repeatedly at scheduled follow-up times, the exact time-to-event may be unknown. Such situations arise when the outcome is classified according to threshold of clinical interest. Then scientific interest is focused on the time at which the threshold is crossed. In these studies recruitment is staggered in time and, increasingly, survival-type methods (Kaplan Meier, 1958; Peto & Peto, 1972, and Cox, 1972) are being pressed into service.

These methods are appropriate for right censored 'time to event data' when the exact time of occurrence is known, but strictly inappropriate when the 'time to event' is known only to lie in an interval. Application of conventional methods to interval 'end' or 'mid' points can lead to bias (Lindsey and Ryan, 1998) and optimistic precision (MacKenzie, 1999). Here, we develop the parametric accelerated life (AL) logistic model (MacKenzie, 1996) in which the baseline hazard follows the time-dependent logistic (TDL) survival model. We compare inference from the correct model with that from

the mis-specified model which uses follow-up times as if they were exact.

2 Likelihood Formulation

Suppose there are $m+1$ scheduled inspection times, $t_o^+, t_1^+, ..., t_m^+$ at which continuous or ordinal responses $Y_0, Y_1, ..., Y_m$ are measured. Let T be a non negative variable denoting the time to some outcome of interest defined on the Ys. Let $S(t;\theta)$ and $\lambda(t;\theta)$ be the corresponding survival and hazards functions, respectively, depending on the unknown vector parameter $\theta \in \Theta$, where $\theta = (\alpha', \gamma', \beta')'$. Then for a sample of n independent subjects it may be shown that the true censored likelihood for the unknown parameters is:

$$L_1(\theta) = \prod_{i=1}^{n} \left\{ S(t_{i_{(k-1)}};\theta) \left[1 - S(t_{i_{k-1}}, t_{i_k};\theta)\right] \right\}^{\delta_i} \left[S(t_i^*;\theta)\right]^{1-\delta_i} \tag{1}$$

where typically n_k patients fail between scheduled examination times $t_{(k-1)}^+$ and t_k^+ for $k = 1, ..., m$ and n_c patients are censored or withdrawn at specific times, t_i^*,such that $n_c + \sum_{k=1}^{m} n_k = n$. Here, $\delta_i = 1$ denotes an event and $\delta_i = 0$ denotes a censored observation. We may compare (1) with the mis-specified censored likelihood resulting from treating the observed inspection times as if they were exact:

$$L_2(\theta) = \prod_{i=1}^{n} \left[\lambda(t_{i_k};\theta) S(t_{i_k};\theta)\right]^{\delta_i} \left[S(t_{i_k}^*;\theta)\right]^{1-\delta_i} \tag{2}$$

Equations (1) and (2) enable us to investigate the effect of mis-specification for any survival model where the function takes closed form. Notice the use of observed inspection times rather than the scheduled times in equation (1).

3 Model Formulation

MacKenzie's (1996) AL logistic survival model is defined by the hazard function

$$\lambda(t;x) \quad = \quad \frac{\lambda \exp(tx'\beta + \gamma)}{1 + \exp(tx'\beta + \gamma)} \tag{3}$$

a form which we have modified to obtain an accelerated hazard model defined by

$$\lambda(t;x) \quad = \quad \frac{\lambda \exp(t\alpha\phi)}{1 + \exp(t\alpha\phi)} \tag{4}$$

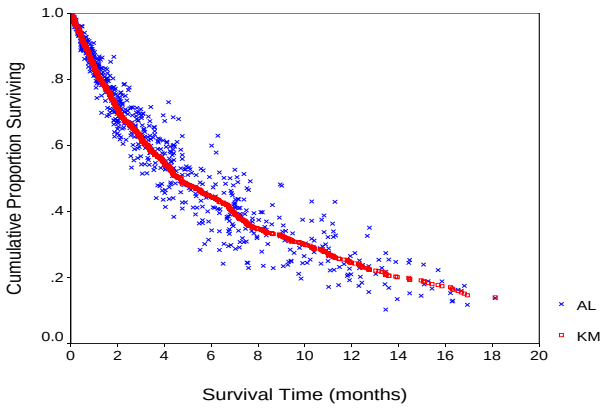


FIGURE 1. *Predicted AL versus KM .*

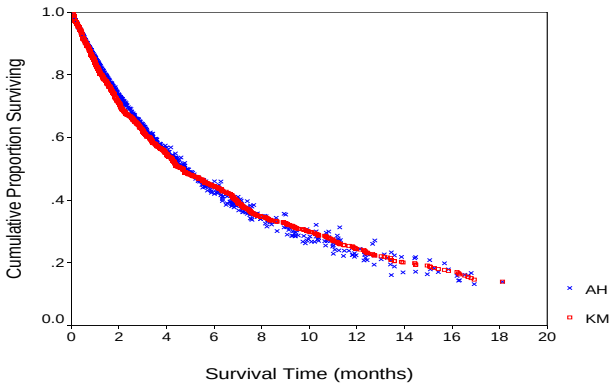


FIGURE 2. *Predicted AH versus KM .*

where $\phi = \exp(x'\beta)$ and we have suppressed the dependence on θ. We compare this model with the corresponding modified accelerated life model defined by

$$\lambda(t|x) = \lambda\phi\frac{\exp(t\alpha\phi)}{1+\exp(t\alpha\phi)} \tag{5}$$

a form which is recognisably different from (4).

TABLE 1. *Comparison of Mis-specified and True Models Estimators: AL Model.*

Mid-point, Regular follow up (3,6,9,12,15,18,21,24)							
		Mis-specified			True		
	n	$\hat{\phi}^*$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\phi}^*$	$\hat{\alpha}$	$\hat{\beta}$
	$\phi^* = -0.6$, $\alpha = 0.2$, $\beta = -0.5$, % within censoring =0						
Mean	100	-1.007	1.901	-0.451	-0.601	0.251	-0.511
(se.)		(0.160)	(0.250)	(0.161)	(0.215)	(0.201)	(0.178)
Mean	500	-1.052	1.976	-0.438	-0.626	0.201	-0.501
(se.)		(0.053)	(0.120)	(0.071)	(0.131)	(0.135)	(0.078)
	$\phi^* = -0.6$, $\alpha = 0.2$, $\beta = -0.5$, % within censoring=30						
Mean	100	-1.346	1.882	-0.443	-1.060	0.371	-0.479
(se.)		(0.175)	(0.294)	(0.188)	(0.207)	(0.306)	(0.205)
Mean	500	-1.403	1.962	-0.442	-1.127	0.531	-0.481
(se.)		(0.061)	(0.168)	(0.089)	(0.151)	(0.385)	(0.091)

TABLE 2. *Comparison of Mis-specified and True Models Estimators: AH Model.*

Mid-point, Regular follow up (3,6,9,12,15,18,21,24)							
		Mis-specified			True		
	n	$\hat{\phi}^*$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\phi}^*$	$\hat{\alpha}$	$\hat{\beta}$
	$\phi^* = -0.6$, $\alpha = 0.2$, $\beta = -0.5$, % within censoring =0						
Mean	100	-1.020	1.618	-0.034	-0.624	0.253	-0.552
(se.)		(0.094)	(0.203)	(0.221)	(0.180)	(0.211)	(1.112)
Mean	500	-1.053	1.742	-0.045	-0.635	0.225	-0.538
(se.)		(0.040)	(0.103)	(0.104)	(0.129)	(0.125)	(0.588)
	$\phi^* = -0.6$, $\alpha = 0.2$, $\beta = -0.5$, % within censoring=30						
Mean	100	-1.024	1.688	-0.024	-0.632	0.248	-0.393
(se.)		(0.088)	(0.211)	(0.218)	(0.197)	(0.199)	(1.206)
Mean	500	-1.055	1.744	-0.045	-0.634	0.212	-0.530
(se.)		(0.040)	(0.108)	(0.099)	(0.125)	(0.109)	(0.584)

4 Simulation Study

The object of the simulation study is to quantify the degree to which inference about the parameters in the AH & AL models, especially β, is compromised by the use of the mis-specified likelihood. We investigate the 2-sample case, mimicking a RCT in which scientific interest is focused on estimating the treatment effect and its associated standard error. The simulation parameters include: sample size, percentage censored, patterns of

follow-up examination is regularly and irregularly spaced, the model parameters (θ). The maximum likelihood estimates will be calculated using the correct and the mis-specified likelihoods.

5 Results

First we compared models (4) and (5) using lung cancer data, and present the conditional fits obtained by each regression model and the marginal fit of the Kaplan Meier estimator. The (AH) model shows a better fit compared with the (AL) model (Figures 1, 2).
Second, we report a subset of the complete simulation using mid-points in the mis-specified likelihood. Tables 1 and 2 show the MLE's for the three parameters using a regular visit schedule. Note that we report $\phi^* = \log_e(\lambda)$ in the tables. Overall, the true likelihood provided consistently better estimates with superiority for the AH model compared with the AL model, when we allowed for drop-out and using a regular schedule. The mis-specified likelihood also produced standard errors which were artificially precise.

6 Summary

The idea of an accelerated hazard model is new. To our knowledge this is the first time that they have been described, and compared empirically with classical accelerated life models, allbeit in the context of a single family of survival models - the GTDL (MacKenzie, 1996). The results of the numerical analysis favour the AH model suggesting that the model may be useful in practice. The advantages of these parametric models stem from the closed forms taken by survivor functions and the fact that when $\beta = 0$ the underlying survival functions have testable parametric forms. We have demonstrated by simulation, the use of these two models in the analysis of interval censored survival data arising in longitudinal randomized controlled trials.

References

Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Association, Series B*, **34**, 187–220.

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457–481.

MacKenzie, G. (1996). Regression models for survival data: the generalized time-dependent logistic family. *The Statistician*, **45**(1), 21–34.

MacKenzie G. (1999). Survival analysis for longitudinal data. *14th International Workshop on Statistical Modelling*. Graz, Austria. pp. 259–264.

Application of General Finite Mixture Models to Reliability Data Using Likelihood Estimation

E. Andries[1], K. Croes[2], L. De Schepper[1] and G. Molenberghs[1]

[1] Limburgs Universitair Centrum, Institute for Materials Research (IMO), Materials Physics Division, Wetenschapspark 1, B-3590 Diepenbeek (Belgium)
[2] XPEQT,Transportstraat 1,B-3980 Tessenderlo (Belgium)

Abstract: Likelihood estimation of the general finite mixture model is considered. A short discussion on this likelihood method is given. The phenomenon of spurious maxima is explained and its relation with sample size. The method is demonstrated on two real reliability data examples.

Keywords: General finite mixture model; Likelihood estimation; Spurious maxima; Reliability data.

1 Introduction

A lot of today's reliability data obtained from experiments with microelectronic components give evidence of bi or even multi modal failure data. Although reliability engineers know mostly whether there is more than one failure mechanism involved, at the end of the experiment it is or too difficult or too expensive to recover the specific failure reason of each device. A large part of these failure data can be modeled by means of a finite mixture. In particular, mixtures with mixing over all parameters are of interest since, due to the nature of much reliability data, a common shape parameter for the component densities cannot a priori be assumed.
A general M-component finite mixture has the following density:

$$f_M(x|\theta) = \sum_{m=1}^{M} \pi_m f(x|\mu_m, \sigma_m) \tag{1}$$

with $\sum_{m=1}^{M} \pi_m = 1$, $f(x|\mu_m, \sigma_m)$ the density of a 2-parameter distribution, μ_m a scale and σ_m a shape parameter. The problem with these mixtures is that a maximum likelihood estimate (MLE), defined as the global maximum of the likelihood, does not exist. However, the likelihood does have a local maximum with, very importantly, good statistical properties.
The aim of this paper is first to discuss shortly this likelihood theory, which is far from new, acknowledged by some authors, but still rarely applied.

Second to tackle the problem of spurious maxima and sample size, and third to demonstrate the method on two sets of reliability data.

2 Likelihood Estimation

It is well known that the likelihood function for a mixture with density (1) is unbounded at some points on the edge of the parameter space. As a result an MLE does not exist. Nevertheless, both empirical and theoretical evidence proved for finite normal mixtures the existence of some local maximum of the likelihood function with good statistical properties, i.e. consistent, asymptotically normal and efficient (Quandt, 1972; Kiefer, 1978). An explanation for this is given by the different conditions determining the existence of a consistent global and a consistent local maximum of the likelihood. Namely, under the conditions of Cramér (1946), the likelihood equations (LEQ) have a (in essence unique) consistent, asymptotically normal and efficient solution which, with probability tending to one as the sample size tends to infinity, corresponds to a local maximum. On the other hand, the different and more demanding conditions of Wald (1949) ensure the consistency of the classical MLE.
While for a mixture with common shape parameter both set of conditions hold, only Cramér's conditions apply for a general M-component mixture . Importantly, whether we either work with a mixture with common or with unequal shape parameters, in essence the same kind of estimate is obtained from the LEQ, in spite of the convention of terminology to only call the first an MLE. The latter will be referred to as a likelihood estimate (LE).
The problem is not entirely solved yet since the likelihood function for (1) has multiple roots and it is not specified which is the proper one. It can be proven that for many general finite mixtures, such as the (log)normal or Weibull mixture, the largest local maximum of the likelihood corresponds to those well-behaved estimates. This gives a criterion similar to ML estimation. But, not everyone agree on this as McLachlan et al. (2000), among others, claim that a spurious maximum could then be chosen as LE.

3 Spurious Maxima

What is meant with a spurious maximum? No unambiguous definition exists yet, but mostly the corresponding mixture is characterized by a small proportion or shape parameter for one of its components. Since a spurious maximum cause problems when it has the largest likelihood value, some authors suggest to first remove all solutions of the LEQ corresponding to such maxima and then to choose among the remaining roots the solution with the largest likelihood as LE. Although these maxima should be considered with care, this procedure is dangerous, highly subjective and we do not recommend it.

TABLE 1. *Some local maxima of the likelihood of two simulated datasets. Estimates in bold correspond to the maximum closest to the true values.*

n	$\mu_1 = 0$	$\sigma_1 = 0.5$	$\mu_2 = 3$	$\sigma_2 = 1$	$\pi_1 = 0.2$	Likelihood
	-1.029	0.00175	2.493	1.470	0.0399	-85.122
50	1.387	1.444	3.627	0.553	0.569	-88.398
	-0.249	**0.694**	**2.996**	**0.987**	**0.198**	**-88.571**
	-0.255	**0.627**	**2.895**	**0.984**	**0.128**	**-200.921**
120	3.246	0.685	1.693	1.543	0.514	-202.120
	2.874	0.0000541	2.485	1.425	0.0166	-202.628

The point is that spurious maxima are not only related to the largest likelihood criterion and the finite mixture case. They exist as soon as Cramér's conditions hold and as the LEQ have multiple roots; irrespective of the fact whether we search for a local or a global maximum. Their appearance as the largest maximum is primarily due to the ambiguity in the statement of a consistent root and related with sample size. Indeed, consistency is a limiting property. As a result an improper estimate can be the outcome of the likelihood or ML method if the sample size n is too small. We define a spurious maximum as each maximum of the likelihood that is not closest to the true values, with *closest* defined by some distance.
How can we then obtain a proper estimate from the LEQ in case of multiple roots? First, use always a consistent procedure (e.g. the largest local or global criterion). Second, choose the sample size large enough. If the latter is not possible, one should take into consideration another method or look whether there is relevant information about the possible true values. A too small sample can often be recognized through the likelihood value of distinct spurious maxima, i.e. maxima for which one of the component densities corresponds to no more than a few data values (if these data values are not clearly separated from the others). Namely, simulations indicated that when n is too small, often at least one of these maxima has a highest likelihood value. As an example, Table 1 gives some local maxima of the likelihood of two simulated datasets from a two-component normal mixture. The second dataset is based on the first but with 70 extra values generated. As seen, for the smallest dataset, the maximum with the highest likelihood value is distinct spurious (π_1 is less than 2/50), while for the larger dataset the proper maximum has the highest likelihood value.

4 Examples

We demonstrate the likelihood estimation method on two datasets obtained from experiments carried out at IMO. The experiments consisted of accelerating the failure mechanisms of a micro-electronic component by means of

TABLE 2. *Some maxima of the likelihood for the datasets of examples 1 and 2. The first maximum of each example is the largest local. Estimated parameters are the mean and shape of the mixture distribution of the log failure times.*

Example	$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{\mu}_2$	$\hat{\sigma}_2$	$\hat{\pi}_1$	Likelihood
	6.164	0.236	7.022	0.251	0.286	-66.100
1	6.745	0.0000355	6.777	0.463	0.0159	-72.545
	6.864	0.0000436	6.775	0.463	0.0158	-72.918
	5.329	0.0000684	5.009	0.429	0.0292	-40.444
2	5.071	0.0000783	5.020	0.432	0.0291	-41.375
	4.086	0.000730	5.041	0.413	0.0286	-43.137
	4.220	0.0518	5.080	0.389	0.105	-44.575
	4.697	0.340	5.374	0.166	0.641	-45.427

increasing certain stress factors (such as temperature, current, ...). Main interest is in the estimation of the failure-time distribution. For the components under study it is known that there could be a second failure mechanism, leading to bimodal failure data.

4.1 Example 1

The failure times of 125 commercial metal film resistors, stressed at a temperature of 165 °C, were measured. Figure 1(left) shows a lognormal QQ-plot of the data. Generally, the failure times for this type of component are lognormally distributed. Since the data suggest two failure modes, a two-component lognormal distribution is estimated. The local maxima of the likelihood function are searched for and the most important ones are indicated in Table 2. As noticed, the largest local maximum is not a distinct spurious maximum and its likelihood value is much larger than the second largest maximum. So, there is no reason to mistrust the largest local maximum. The fitted distribution is shown in Figure 1(left). One can now proceed as in case of ML estimation and carry out tests, construct confidence intervals, ... in the usual way.

4.2 Example 2

Interconnects were stressed at 80 °C and 0.75MA/cm^2. All 68 devices under test failed. A Weibull QQ-plot of the failure times is shown in Figure 1(right). Previous experiments indicated that there could be two failure mechanisms. So, a two-component Weibull mixture is used to fit the data. The likelihood is scanned for local maxima and some of them are tabulated in Table 2. In contrast to the first example, the largest local maximum is

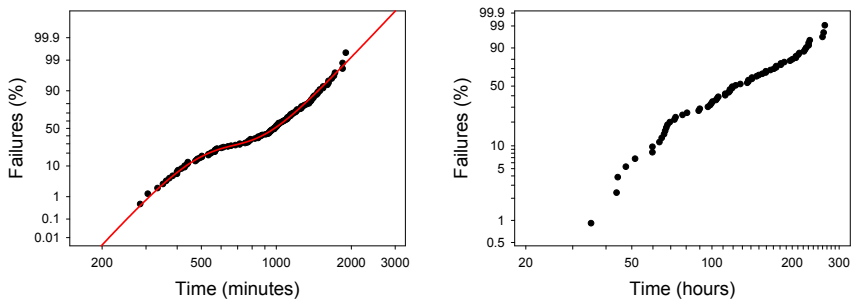


FIGURE 1. *(left) Lognormal QQ-plot ex.1; (right) Weibull QQ-plot ex.2.*

now distinct spurious. Although the last two maxima in the table correspond to reasonable estimates, it is dangerous to choose one of these two as the LE. Indeed, depending on the chosen maximum other inference results are obtained, which could lead to wrong reliability predictions and conclusions. If there are truly two failure modes, this is not clearly seen yet. Consequently, more data is needed or other techniques have to be applied.

5 Conclusions

Despite the nonexistence of the MLE for general finite mixtures, there exists a root of the LEQ with good statistical properties. It is the same kind of estimate as the MLE, called the LE and corresponds for a lot of cases to the largest local maximum of the likelihood.
The appearance of spurious maxima is inherently linked to the presence of multiple roots in the LEQ and independent of the fact whether one search for the largest local or global maximum. When the likelihood function is dominated by distinct spurious maxima, the sample is most likely too small and none of the roots of the LEQ can be trusted.

Acknowledgments: This work was supported by the Flemish Science foundation (IWT).

References

Quandt, R.E. (1972). A new approach to estimating switching regressions. *Journal of the American statistical association*, **76**, 306–310.

Kiefer, N.M. (1978). Discrete parameter variation: efficient estimation of a switching regression model. *Econometrica*. **46**(2), 427–433.

Cramér, H. (1946). *Mathematical methods of statistics (ch. 33)*. Princeton University Press.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, **20**, 595–601.

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

A Comparison of Non-PH and PH Gamma Frailty Models

Milica Blagojevic[1], Gilbert MacKenzie[1], and Il Do Ha[2]

[1] Centre for Medical Statistics, Keele University, Keele, Staffordshire ST5 5BG, UK. Email: m.blagojevic@maths.keele.ac.uk, g.mackenzie@keele.ac.uk
[2] Faculty of Information Science, Kyungsan University, Kyungsan, 712-240, South Korea. Email: idha@kyungsan.ac.kr

Abstract: The Non-PH Canonical Time Dependent Logistic survival regression model described by MacKenzie (1996, 2002), is extended by incorporating a multiplicative Gamma frailty component into the hazard function. The resulting model is obtained in closed form and its properties are compared with the classical PH, Weibull frailty regression model described by (Hougaard, 1994). The performance of the models, with and without frailty, is investigated by re-analyzing some data from the NI lung cancer study (MacKenzie, 1996).

Keywords: PH and non-PH models; Gamma frailty; Canonical logistic.

1 Introduction

The Weibull proportional hazards (PH) regression survival model has been extended to a frailty model by means of a multiplicative random effect acting on the hazard function (Hougaard, 1994). Classically the random component is assumed to follow a Gamma distribution, which is mathematically tractable and leads, after marginalization, to a closed form for the resulting frailty distribution. However, not all survival data are PH and it is therefore useful to have alternative non-PH models. This is relevant as, increasingly, random effect models are being used to analyze multivariate survival data (Ha, Lee and Song, 2001, Ha and Lee, 2003).
A flexible non-PH model is the Canonical Time-Dependent Logistic (CTDL) described by MacKenzie (1996, 2002). We generalize this model by including a multiplicative Gamma frailty term in the hazard function. The resulting frailty model is obtained in closed form and we compare its properties with the Weibull frailty model, noting the connection with a general class of frailty models described by Aalen (1988). Moreover, we investigate the performance of the four models, Weibull and CTDL with and without frailty, using data from the Northern Ireland lung cancer study.

2 Parametric Regression Models with Frailty

Consider a basic survival regression model with failure time density $f(t|\theta,\beta)$, hazard function $\lambda(t|.)$ and survivor function $S(t|.)$, where typically θ is a vector valued parameter and β is a regression parameter. Assume that a random variable U, with density $g(u|\sigma^2)$, denotes the unobservable individual (i.i.d.) frailties and that $E(U) = 1$ and $V(U) = \sigma^2$. Then, given data (t_i, x_i, δ_i) for $i = 1 \ldots n$ subjects, a target vehicle for inference is the marginal likelihood of the parameters of interest

$$L_f(\theta,\beta,\sigma^2) = \prod_{i=1}^{n} \int_0^\infty \lambda(t_i|u_i,\theta,\beta)^{\delta_i} S(t_i|u_i,\theta,\beta) g(u_i|\sigma^2) du_i \tag{1}$$

where δ_i is the censoring indicator and f denotes the marginal *frailty survival model*, derived from $f(t|u,\cdot)$ using the *frailty distribution* density $g(u|\cdot)$.
In more general cases, the marginal likelihood may be analytically intractable, when recourse to numerical methods of integration may be required. Alternatively, the h-likelihood method, extended to survival models by Ha, Lee and Song (2001), has the obvious advantage of dispensing with the need for marginalization in several important classes of statistical models. Here, we adopt a classical approach to the derivation of $L_f(\cdot)$ for two parametric survival models with Gamma frailty and obtain the resulting marginal frailty survival models in closed form.

2.1 CTDL Model

A non-PH model, the CTDL regression model (MacKenzie, 1996), is defined by the hazard function

$$\lambda(t|x) = \lambda p(t|x), \tag{2}$$

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 $\theta' = (\lambda, \alpha, \beta)$.
The corresponding survival function is

$$S(t|x) = \{(1 + \exp(t\alpha + x'\beta))/(1 + \exp(x'\beta))\}^{-\frac{\lambda}{\alpha}} \tag{3}$$

whence the censored log-likelihood becomes

$$\ell(\lambda,\alpha,\beta) = \sum_{i=1}^{n} \left[\delta_i \log_e \lambda + \delta_i \log_e p_i + \frac{\lambda}{\alpha} [\log_e g_i + \log_e g_i] \right] \tag{4}$$

where,

$$\begin{aligned} p_i &= \exp(t_i\alpha + x_i'\beta)/\{1+\exp(t_i\alpha + x_i'\beta)\} \\ q_i &= 1/\{1+\exp(t_i\alpha + x_i'\beta)\} \\ g_i &= 1+\exp(x_i'\beta) \end{aligned} \tag{5}$$

and where, for notational convenience, we have suppressed the dependence on time and the covariates on the LHS of (5).
When developing the CTDL-gamma mixture model, we assume that the random component has a multiplicative effect on the hazard, such that $\lambda(t;x,u) = u\lambda(t;x)$. If U follows a Gamma distribution with $E(u) = 1$ and $V(u) = \sigma^2$ then $g(u|\sigma^2) = u^{\frac{1}{\sigma^2}-1}\exp(-u/\sigma^2)/\Gamma(1/\sigma^2)\sigma^{2\frac{1}{\sigma^2}}$. We may then integrate out the random effect to obtain the survivor function for the resulting frailty survival distribution, viz

$$S_f(t|x) = \int_o^\infty S(t|x,u)g(u|\sigma^2)du \tag{6}$$

$$= \left\{1 - \frac{\lambda\sigma^2}{\alpha}\log_e(q_i g_i)\right\}^{-\frac{1}{\sigma^2}} \tag{7}$$

whence it follows that

$$\lambda_f(t|x) = \lambda p_i / \left\{1 - \frac{\lambda\sigma^2}{\alpha}\log_e(q_i g_i)\right\} \tag{8}$$

results, which enable the censored log-likelihood to be constructed

$$\ell(\lambda,\alpha,\beta,\sigma^2) = \sum_{i=1}^n \left\{\delta_i \log_e(p_i\lambda) - (\delta_i + \frac{1}{\sigma^2})\log_e(1 - \frac{\lambda\sigma^2}{\alpha}\log_e(q_i g_i))\right\} \tag{9}$$

2.2 Weibull Model

The familiar Weibull regression distribution has the following hazard and survival function

$$\lambda(t|x) = \lambda\rho(t\lambda)^{\rho-1}\exp(x'\beta) \tag{10}$$

$$S(t|x) = \exp(-(t\lambda)^\rho e^{x'\beta}) \tag{11}$$

respectively, giving rise to the censored log-likelihood

$$\ell(\lambda,\rho,\beta) = \sum_{i=1}^n \left[\delta_i \log_e(\rho\lambda^\rho t_i^{\rho-1} e^{x_i'\beta}) - (\lambda t)^\rho e^{x_i'\beta}\right] \tag{12}$$

In deriving the marginal frailty distribution, we make the same assumptions and use the same method as in section 2.1. We find that

$$S_f(t|x) = \left\{1 + \sigma^2(t\lambda)^\rho e^{x'\beta}\right\}^{-\frac{1}{\sigma^2}} \tag{13}$$

and

$$\lambda_f(t|x) = \lambda(t\lambda)^{\rho-1}\rho e^{x'\beta}/(1+\sigma^2(t\lambda)^{\rho}e^{x'\beta}) \tag{14}$$

yielding the log-likelihood

$$\ell(\lambda,\rho,\beta,\sigma^2) = \sum_{i=1}^{n}\left\{\delta_i \log_{\mathrm{e}}(\lambda^{\rho}t_i^{\rho-1}\rho e^{x_i'\beta}) - (\delta_i+\tfrac{1}{\sigma^2})\log_{\mathrm{e}}(\sigma^2(t_i\lambda)^{\rho}e^{x_i'\beta}+1)\right\}$$

3 Example Data Analysis

The data analyzed form part of a population-based prospective study of incident cases of lung cancer diagnosed in Northern Ireland in one year. This multi-source study identified 900 incident cases in which outcome was missing in 25 and a further 20 were diagnosed at post-mortem. We analyzed 'Time from Diagnosis to Death or Censoring' in relation to a range of covariates, but, to illustrate the models, we present a detailed analysis of two covariates, age at diagnosis and sex of the patient. The model fitting procedure was implemented in S-Plus (V4.5) and in R.

The results are presented in Table 1. The analysis illustrates some important findings. First the significant and adverse effect of Age is statistically significant in all four models fitted, although the magnitude of the estimated effect and the corresponding standard error varies. The non-significance of the Sex effect is also confirmed in all models. In the CTDL model $\hat{\alpha}$ is statistically significant and negative so that the trend in the hazard is decreasing with time - which, potentially, is a frailty signature. When Gamma frailty is added to the CTDL $\hat{\alpha}$ becomes non-significant, suggesting that the significant negative trend resulted, at least in part, from heterogeneity. The standard errors in the frailty models are all increased suggesting that the non-frailty models under-estimate the dispersion in the data.

Likelihood ratio tests were conducted within model family to test the absence of the frailty component, i.e. $H_0 : \sigma^2 = 0$. Note that such a hypothesis is on the boundary of the parameter space, so the critical value is $\chi^2_{2\lambda}$ for a size λ test (Chernoff, 1954; Vu and Knuiman, 2002).For the CTDL family, the difference $-2(\hat{\ell} - \hat{\ell_f})$ is 4.24 and for the Weibull family it is 21.68, whence the null hypothesis is rejected for both models by a 5% level ($\chi^2_{1,0.10} = 2.71$). Thus, the addition of a frailty component is justified, especially in the Weibull family. Moreover, as judged by the usual AIC criterion, the Weibull-frailty model is the best model.

However, inspection of the fitted models (not shown) reveals that the AIC is misleading. The CTDL model without frailty is superior to the Weibull model without frailty and the best fit, overall, is provided by the CTDL frailty model. These findings persist when the data are categorized by estimated risk score ($x'\hat{\beta}$) into quartiles and compared with the corresponding conditional Kaplan-Meier estimators.

TABLE 1. *Maximum Likelihood Estimates & (s.e.): four models, with Age (β_1) and Sex (β_2) fitted ($\hat{\lambda}_c = \lambda$ in the CTDL, $\hat{\lambda}_w = \lambda$ in the Weibull, GF = Gamma Frailty).*

Parameter	CTDL	Weibull	CTDL+GF	Weibull+GF
$\hat{\alpha}$	-0.1165	-	-0.0342	-
	(0.0239)	-	(0.0770)	-
$\hat{\lambda_c}$	+0.2148	-	+0.2169	-
	(0.0352)	-	(0.0373)	-
$\hat{\lambda_w}$	-	+0.0309	-	+0.0308
	-	(0.0130)	-	(0.0145)
$\hat{\rho}$	-	+0.8640	-	+0.8186
	-	(0.0280)	-	(0.2181)
$\hat{\beta}_1$	+0.0140	+0.0170	+0.0206	+0.0302
	(0.0071)	(0.0050)	(0.0114)	(0.0082)
$\hat{\beta}_2$	+0.0177	+0.0175	-0.0673	+0.0250
	(0.2000)	(0.0820)	(0.4782)	(0.1197)
$\hat{\sigma}^2$	-	-	+0.4402	+1.1194
	-	-	(0.2054)	(0.0751)
$\hat{\ell}$	-2053.299	-2052.399	-2051.166	-2041.556

4 Final Remarks

In this paper we have derived a new non-PH based Gamma frailty model and compared it with the standard PH-based Gamma frailty competitor. In the data analyzed the interpretation of the fixed effects was similar, but the fit of the CTDL Gamma frailty model was demonstrably superior and it is therefore a more natural vehicle for inference, among the frailty models considered. Of course, this will not always be the case, but the new model is flexible and clearly provides a viable alternative to the PH models considered. Further work is required on discriminating between these survival models and on developing suitable measures of Goodness of Fit for non-nested models with frailty in different model classes, when the AIC may be misleading.

References

Aalen, O.O. (1988). Heterogeneity in Survival Analysis. *Statistics in Medicine*, **7**, 1121-1137 .

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics*, **25**, 573-578.

Ha, I. D., Lee, Y., and Song, J.-K. (2001). Hierarchical likelihood approach for frailty models. *Biometrika*, **88**, 233-243.

Ha, I. D. and Lee, Y. (2003). Estimating frailty models via Poisson hierarchical generalized linear models. *Journal of Computational and Graphical Statistics*. In press.

Hougaard, P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika*, **71**, 75–83.

Hougaard, P. (1994). Heterogeneity models of disease susceptibility, with applications to diabetic nephropathy. *Biometrics*, **50**, 1178–1188.

MacKenzie, G. (1996). Regression models for survival data: the generalised time dependent logistic family. *Journal of the Royal Statistical Society, 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.

MacKenzie, G. (2002). A logistic regression model for survival analysis. *Proceedings of the 17th IWSM, Chania, Crete*, 105–113.

Vu, H. T. V. and Knuiman, M. W. (2002) A hybrid ML-EM algorithm for calculation of maximum likelihood estimates in semiparametric shared frailty models. *Computational Statistics and Data Analysis*, **40**, 173–187.

Association between Air Pollution and Health. Statistical Analysis of a Longitudinal Study with a Binary Outcome

Susanne Breitner[1], Annette Peters[2], Helmut Küchenhoff[1], Angela Ibald-Mulli[2], and H.-Erich Wichmann[2]

[1] Department of Statistics, Ludwig-Maximilians-Universität München, Akademiestr. 1, D-80799 München
[2] GSF, National Research Center for Environment and Health, Institute of Epidemiology, D-85758 Neuherberg

Abstract: We present two different logistic regression models for longitudinal air pollution data. After the confounder modelling we first use marginal regression models to take account of the autocorrelation structure of the data. As a second approach we then present Bayesian generalized additive mixed models with an interaction between the trend and a patient-specific random effect to account for unobserved heterogeneity and autocorrelation. We apply this methods to data of the project EPA STAR.

Keywords: Longitudinal data; Logistic regression; Marginal regression models (GEE); Bayesian generalized additive mixed models (GAMM).

1 Introduction

One objective of the project EPA STAR (**E**nvironmental **P**rotection **A**gency; **S**cience **T**o **A**chieve **R**esults: Inflammatory Response and cardiovascular risk factors in elderly subjects with angina pectoris or COPD in association with fine and ultrafine particles) is to characterize the association between ambient particle exposures and changes in biomarkers of inflammation in the airways and the blood of patients with stable coronary artery disease (CAD). Therefore 60 male non-smokers, aged between 50 and 80 years, were recruited from local practitioners. Further participants have physician diagnosed coronary artery disease or stable angina pectoris or take angina pectoris medication.
Among other things the panelists recorded cardiovascular symptoms and medication intake daily over a period of about six months. Ambient air pollutants as well as meteorological parameters were measured at local monitoring stations on a daily basis.
The main statistical challenge of this type of observational study is to find the effect of air pollution on the symptoms in the presence of confounders, autocorrelation and random effects.

2 Modelling

To analyze the association between air pollution and health outcomes we use logistic regression models for longitudinal data controlling for confounders and autocorrelation.
Responses $y_{it}, i = 1, \ldots, n,\ t = 1, \ldots, T_i$ are binary, with $y_{it} = 1$ for the presence and $y_{it} = 0$ for the absence of a symptom. We assume, that the probability of appearance of a symptom follows a logit model

$$\begin{aligned} \text{logit}(E(y_{it}|z_{it})) &= \log \frac{P(\text{appearance})}{P(\text{no appearance})} \\ &= \eta_{it} = \alpha + \beta x_{t-l} + \gamma_1 z_{it1} + \ldots + \gamma_p z_{itp} \\ &\quad + g_1(v_{t1}) + \ldots + g_k(v_{tk}). \end{aligned}$$

x_{t-l} denotes the (lagged) measure of the air pollutant, z_{itj} are confounder variables with a linear effect and v_{ts} confounders with nonlinear effects.

The modelling principle has two steps: First, the confounder model is fitted. A step-wise model selection procedure in S-PLUS (2001) is used to determine the optimal confounder model. As confounder variables are considered: an indicator variable for each subject, long-term time trend, medication intake, influenza, temperature, relative humidity and barometric pressure each with lag 0 to lag 5 (a lag is the assumed time period between exposure and effect) and an indicator variable for weekday versus weekend. For each metrical covariate both linear and non-linear terms are allowed. The step-wise procedure selects whether each covariate should be included, and if so, whether the metrical covariate should be linear or non-linear. The Akaike Information Criterion (AIC) is used in this algorithm for variable selection.
Due to the problems discussed in Dominici et. al. (2002) we use more stringent convergence parameters than the default settings in S-PLUS for all confounder models. However, at this stage, the procedure does not take any autocorrelation structure of the data into account because this would result in a quite complex statistical computation.

2.1 Marginal Regression

To allow for autocorrelation of the response we then use Marginal regression models for longitudinal data. The modelling and estimation approach is based on generalized estimating equations (GEE). In these models, the effect of covariates on responses and the association between responses is modelled separately.
To be able to fit the GEE model the non-linear and polynomial confounder components are transferred to SAS (2001) with the same degrees of freedom. The impact of calendar time and the possibly non-linear effects of

meteorology are modelled with regression splines or polynoms. By using parametric smoothers we avoid the problem of concurvity (Ramsay et. al. (2003)). The effects of the remaining confounders such as medication intake and of the air pollutant are considered as linear. Additionally a fixed patient-specific effect is included.

2.2 Bayesian GAMM

As a second model approach, we use Bayesian generalized additive mixed models (Bayesian GAMM). This methods allow for a predictor with semiparametric additive form. We now model the impact of the global trend and the possibly non-linear effects of meteorology nonparametrically using P-splines with second order random walk models as priors for the smooth functions. The effects of the remaining covariates, especially of air pollutants are considered as fixed. For the fixed effect parameters we assume independent diffuse priors. Furthermore, interactions between the trend and a patient-specific random effect account for unobserved heterogeneity and autocorrelation. For this interactions we make the assumption that the parameters are i.i.d. Gaussian.
Additionally we assume highly dispersed inverse gamma priors for variances in a further stage of hierarchy. This allows for simultaneous estimation of the unknown function and the amount of smoothness.
Inference is fully Bayesian via Markov Chain Monte Carlo (MCMC) techniques. All Bayesian analyses are performed with BayesX (2000).

3 Results

Figure 1 displays the calendar time trend and a parametric estimate for relative humidity together with errors bands provided by S-PLUS. The trend is slowly declining over the study period. The effect of relative humidity is clearly non-linear. The curve indicates that not only low relative humidity but also relative humidity between 80 and 95% increases the risk.
Table 1 gives the estimates with standard errors and p values for the remaining confounder effects. It is seen that medication intake increases the

TABLE 1. *Estimates of constant confounder parameters for the symptom chest pain.*

Covariate	Coefficient	Std. Error	p value
Intercept	-3.100073	0.575742	<0.000001
Temperature, lag 5	-0.022897	0.012297	0.062678
Medication intake	1.889101	0.282345	<0.000001

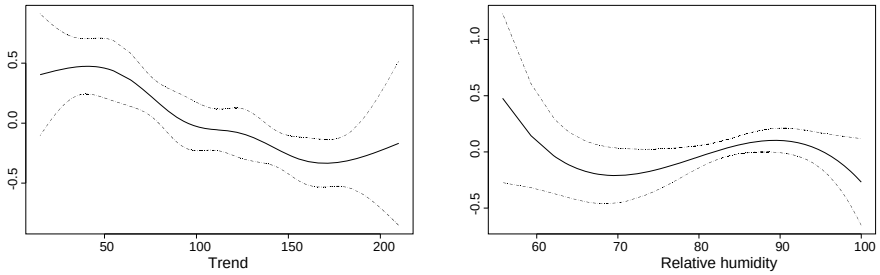


FIGURE 1. *Estimated trend and estimated effect of relative humidity together with error bands provided by S-PLUS.*

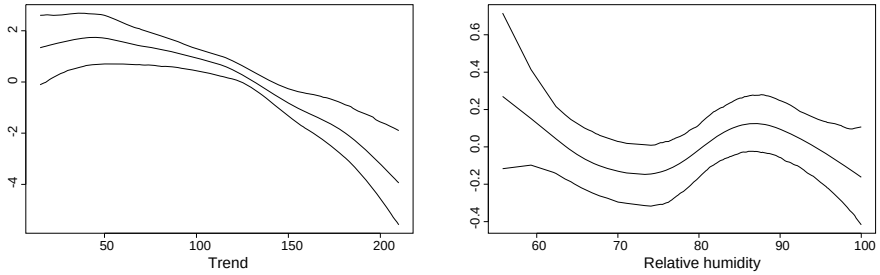


FIGURE 2. *Estimated trend and estimated effect of relative humidity. Shown is the posterior mean within 90% pointwise credible intervals.*

TABLE 2. *Bayesian estimates of constant confounder parameters for chest pain.*

Covariate	Mean	10% quantil	90% quantil
Intercept	-0.8546	-2.2439	0.9403
Temperature, lag 5	-0.0258	-0.0430	0.0095
Medication intake	1.1114	0.6121	1.5952

risk while an increasing temperature with lag 5 decreases the probability of chest pain.

The Bayesian estimation results of the nonparamtric confounder terms and the trend are shown in Figure 2. Corresponding to the parametric estimates

TABLE 3. *Estimated effects of NO_2 and ultrafine particles (GEE).*

Lag	Odds Ratio/IQR	95% CI	p value	Odds Ratio/IQR	95% CI	p value
0	0.940	(0.984,1.076)	0.371	1.012	(0.888,1.152)	0.862
1	1.027	(0.991,1.177)	0.701	0.945	(0.816,1.095)	0.452
2	1.230	(1.001,1.487)	0.032	1.121	(0.922,1.364)	0.252
3	1.038	(0.991,1.214)	0.636	1.061	(0.907,1.242)	0.461
4	0.981	(0.981,1.220)	0.865	0.970	(0.821,1.147)	0.723
5	0.970	(0.982,1.179)	0.759	0.895	(0.705,1.137)	0.364

TABLE 4. *Estimated effects of NO_2 and ultrafine particles (Bayesian GAMM).*

Lag	Odds Ratio	90% Credible Interval	Odds Ratio	90% Credible Interval
0	0.925	(0.839,1.025)	0.984	(0.884,1.104)
1	1.030	(0.918,1.152)	0.962	(0.859,1.069)
2	1.212	(1.086,1.351)	1.096	(0.979,1.223)
3	1.113	(0.994,1.248)	1.086	(0.972,1.212)
4	1.021	(0.918,1.138)	0.998	(0.895,1.112)
5	0.968	(0.871,1.082)	0.907	(0.809,1.010)

the calendar time trend is again declining but with an stronger effect. The curve for relative humidity indicates that low relative humidity and relative humidity between 80 and about 95% increases the risk. This is in agreement to the results shown in Figure 1.
Table 2 gives the posterior means together with 90% credible intervals of the remaining confounder effects. As can be seen the medication intake again increases the risk, while an increasing temperature with lag 5 decreases the probability of chest pain. In comparison to the results in Table 1 the estimations of temperature and medication intake have the same signs, but the effect of medication intake here is less strong.

The Marginal regression results for the gaseous pollutant NO_2 and for particle number concentrations of ultrafine particles by the interquartile range are given in Table 3, showing the strongest associations with a lag of 2 days. For NO_2 this association is even significant.
Table 4 shows the results of the Bayesian pollutant analysis. The strongest associations are again seen with a lag of 2 days whereas for NO_2 this association is even significant. This is in close agreement to the marginal regression approach.

4 Discussion

In this analysis we use two different logistic regression models for longitudinal data controlling for trend, medication intake, influenza, meteorology and autocorrelation. By using parametric regression splines we furthermore avoid the problem of concurvity as discussed in Ramsay et al. (2003). Software has been developed on all aspects of our models, but there are still some technical difficulties to deal with the problems of confounder modelling and autocorrelation simultaneously in one model. In comparison with the Marginal regression model the results of the Bayesian approach generally show similar effects. The main advantage of the hierarchical Bayesian model is the modular structure and flexibility.

References

de Hartog J.J., Hoek G., Peters A., Timonen K.L., Ibald-Mulli A., et al (2003). Effects of fine and ultrafine particles on cardiorespiratory symptoms in elderly subjects with coronary heart disease: the ULTRA study. *American Journal of Epidemiology*, **157**, 613–623.

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

Fahrmeir, L. and Lang, S. (2001). Baysian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society, Series C*, **50**, 201–220.

Lang, S. and Brezger, A. (2000). BayesX - Software for Bayesian Inference based on Markov Chain Monte Carlo simulation techniques. Discussion Paper 187. Universität München.

Ramsey T.O., Burnett R.T. and Krewski, D. (2003). The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. *Epidemiology*, **14**, 18–23.

Spatial Mixture Models for Ordinal Responses: Grazing Impacts in Scotland, UK

Mark J. Brewer[1], David A. Elston[1], and Andrew J. Nolan[2]

[1] Biomathematics and Statistics Scotland, Macaulay Institute, Craigiebuckler, Aberdeen, Scotland, AB15 8QH
[2] Macaulay Institute, Craigiebuckler, Aberdeen, Scotland, AB15 8QH

Abstract: We apply a spatial mixture smoothing model to data arising from a study into grazing and trampling impacts by large herbivores in part of Scotland. The mixture model is shown to allow the separation of larger-scale trend in impacts from more localised variation.

Keywords: Spatial smoothing; Mixture model; Grazing animals.

1 Introduction

We consider two spatial models for ordinal responses, and apply both to data arising from a study of grazing and trampling impacts in the South Loch Tay area of Scotland (Stolte et al, 2003). The first includes an L2-norm spatial random effect as in Besag et al (1991), whereas the second uses mixing weights to combine both L2- and L1-norm spatial terms (see Lawson and Clark, 2002).

Since 1997, the Macaulay Institute has been conducting surveys on the impact of grazing and trampling by large herbivores on areas of upland Scotland, in collaboration with the Deer Commission for Scotland, Scottish Natural Heritage and Deer Management Groups. These surveys are intended to inform decisions relating to the management of grazing uplands and to aid understanding of the effects of different policies. The data in this paper arise from one such study, on the South Loch Tay Deer Management Group area in Perthshire, Scotland. The response is ordinal on a 5-point scale, representing assessments of the intensity of impacts by grazing animals. The classes are Light, Light/Moderate, Moderate, Moderate/Heavy and Heavy. These assessments are made for parcels of land called *part-polygons*, which represent the intersection of contiguous areas of common vegetation type (or *habitats*) and 0.25 km^2 squares of the National Grid.

The habitat of each part-polygon is available as covariate information, as is the estate to which each belongs; there are 28 habitats and 6 estates in South Loch Tay. We model the spatial effects at the 0.25 km^2 grid square level, since some of the part-polygons are too small to be genuinely representative—less than $10m^2$ in a few cases.

2 Spatial Models for Ordinal Responses

We adopt the approach to modelling ordinal response data proposed by Albert and Chib (1993), whereby a latent variable is used to model an underlying unobservable continuous response and where cut-points separate the observed classes. As in Albert and Chib (1993), we use Markov chain Monte Carlo (MCMC) as our main analytic tool.
The density curve specific to each part-polygon is assumed to be Normal with mean μ_k and common variance 1, where $k = 1, 2, \ldots, n$ indexes each individual part polygon, being of habitat h_k and belonging to estate e_k. The cut-points on this continuous (latent) scale are represented by θ_1 to θ_4, and if we further define θ_0, θ_5 to be $-\infty, +\infty$ respectively, then we can define the probability of part-polygon k exhibiting class c grazing impact as

$$\Pr\left(Y_k = c \mid \boldsymbol{\theta}, \mu_k\right) = \Phi\left(\theta_c - \mu_k\right) - \Phi\left(\theta_{c-1} - \mu_k\right) \tag{1}$$

for $c = 1, \ldots, 5$, and where Φ is the standard Normal distribution function. The variance is fixed since the probabilities at (1) are scale-invariant, and we use the value 1 w.l.o.g. here. We consider two spatial models. Both include an L2 spatial effect δ_{g_k}; if we define neigh(g_k) to be the set of n_{g_k} neighbours of grid square g_k then the (implicit) prior specification for δ_{g_k} is

$$\delta_{g_k} |\, \delta_{j \neq g_k} \sim N\left(\bar{\delta}_{g_k}, \frac{\sigma^2}{n_{g_k}}\right) \qquad \text{where} \qquad \bar{\delta}_{g_k} = \sum_{j \in \text{ neigh}(g_k)} \delta_j / n_{g_k}$$

and σ^2 is the random effect variance. The second model also includes an L1 spatial random effect η_{g_k}, and where the product of the priors for $\{\eta_{g_k}\}$ is

$$\lambda^{-m/2} \exp\left\{(2\lambda)^{-1} \sum_{g_k} \sum_{j \in \text{ neigh}(g_k)} |\eta_j - \eta_{g_k}|\right\}.$$

Note that Lawson and Clark (2002) use $\lambda^{-1/2}$ rather than $\lambda^{-m/2}$. The δ_{g_k} and the η_{g_k} are both scaled to have mean 0. The stated aim in Lawson and Clark (2002) is to allow δ_{g_k} to reveal smoothly-changing patterns while η_{g_k} uncovers "discrete jumps". The linear predictor for our first model is

$$\mu_k = \alpha_{h_k} + \beta_{e_k} + \gamma_{h_k, e_k} + \delta_{g_k}$$

where α_{h_k} represents the habitat (random) effect for habitat h_k, β_{e_k} the estate (fixed) effect for e_k, γ_{h_k, e_k} the interaction effect for the habitat/estate combination (random for habitat within estate), and δ_{g_k} the spatial effect for grid square g_k. For the second model, we weight the spatial components by p_k:

$$\mu_k = \alpha_{h_k} + \beta_{e_k} + \gamma_{h_k, e_k} + p_k \delta_{g_k} + (1 - p_k)\, \eta_{g_k}.$$

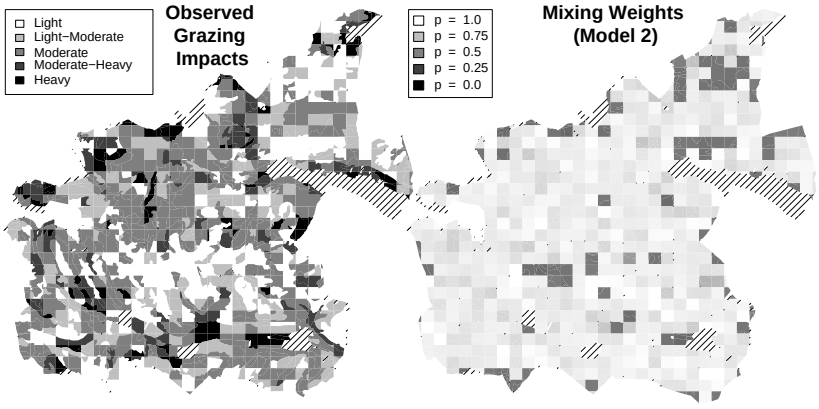


FIGURE 1. *Maps of observed grazing impacts and mixing weights for model 2.*

Finally, we use: weakly informative Gamma priors for precisions and for λ; and diffuse Normal priors for α_{h_k}, β_{e_k}, γ_{h_k,e_k} and θ_c, the latter with constraints to ensure $\theta_{c-1} < \theta_c$ for $c = 0, \ldots, 5$. Fuller details of the model can be found in Brewer et al (2003).

3 Results and Discussion

In this paper we concern ourselves with the spatial smoothing of the two models, and not with the importance or otherwise of the remaining covariates—again, see Brewer et al (2003). In Figure 1 we see the observed grazing impacts. The impacts are not smooth, but we can make out, e.g., areas of lower impacts in the centre and the top right (north-east)—these correspond to an area of high altitude and an area of low animal stocking density respectively. These areas are shown far more clearly in the maps of fitted impacts from both models in Figure 2. The fitted maps are similar, but the model 1 map is smoother, and the model 2 map has more part-polygons in the extreme classes—most noticeably for the Heavy class. Note that there are covariate terms in the models, so we should not expect the maps to be too smooth.

Figure 3 shows the L2 spatial effects δ; there are clear differences here between the two models. The map for model 2 is far smoother than that for model 1—the L1 term η in model 2 has removed the discontinuities, leaving δ to describe the larger-scale smooth spatial trend.

In contrast to the findings in Lawson and Clark (2002), we find more information as to the location of jumps in the mixing weights p than in the η themselves. Hence, we show a map of p from model 2 in Figure 1, where the darker areas relate to greater weight on η, i.e. large jumps in the spatial pattern. Comparing the two maps of Figure 1, it is clear that the weights do pick out areas where the grazing impacts differ most between neighbouring

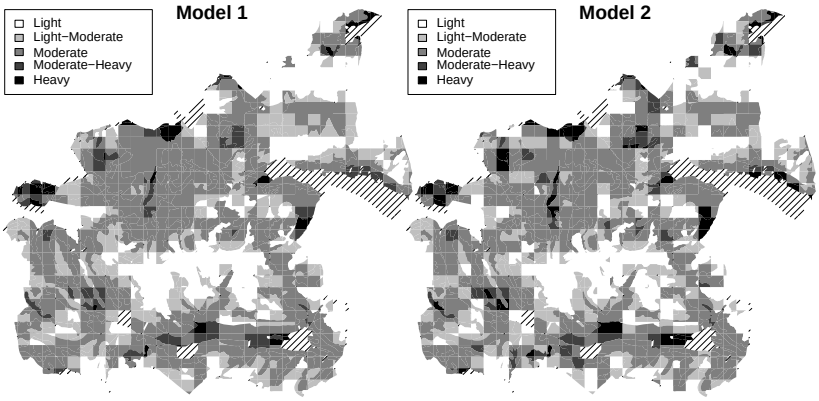


FIGURE 2. *Maps of fitted impacts for models 1 and 2.*

TABLE 1. *DIC values for both models.*

Model	$\bar{D}$	$D(\bar{\mu})$	$p_D(\bar{\mu})$	DIC($\bar{\mu}$)	$D(\bar{\pi})$	$p_D(\bar{\pi})$	DIC($\bar{\pi}$)
1	4993.0	4554.4	438.6	5431.6	4662.7	330.3	5323.3
2	4743.1	4192.5	550.6	5293.7	4310.3	432.8	5176.0

grid squares. Typically, low p values correspond either to areas with low animal numbers (due to altitude, low stocking, etc) or high animal numbers (due to the likely placement of food), where the numbers are not fully explained by habitat or estate. Space limitations prevent us showing the map of η, which in any case appears fairly random due to the high values of p for many of the grid squares (cf. the map of p in Lawson and Clark, 2002).

Finally, we show DIC values (with standardising factor set to 1, see Spiegelhalter et al, 2002) for both models in Table 1. Whether we use the means μ or the class probabilities π to calculate DIC, we see that model 2 is favoured. The standard deviations of the $\bar{D}$ quantities were around 30, so the differences in DIC would appear to be meaningful. Model 2 is far more complex (by the higher p_D values), but this appears to be outweighed by model fit. Also, we note the large differences between the parameterisations—following the advice in Spiegelhalter et al (2003) regarding Normality of likelihoods, we prefer DIC($\bar{\mu}$).

In conclusion, the mixture model with two kinds of spatial random effect provides a better fit to our grazing impact data than a simpler version, allowing a clear separation of overall trend and more localised jumps.

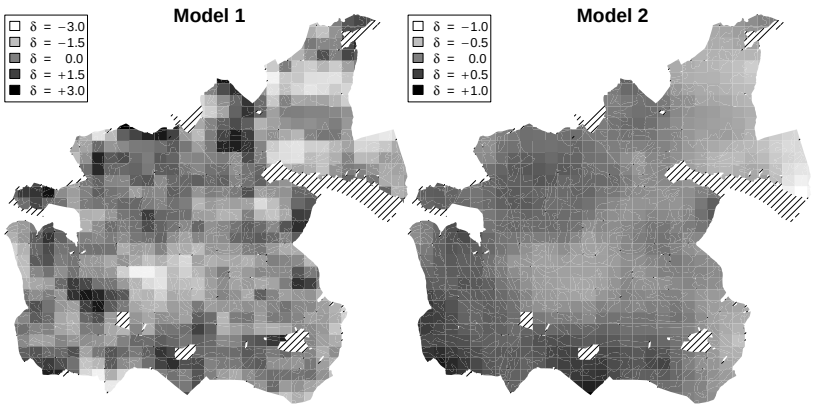


FIGURE 3. *Maps of estimated spatial random effects (δ) for models 1 and 2.*

References

Albert, J.H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**, 669–679.

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

Brewer, M.J., Elston, D.A., Hodgson, M.E.A., Stolte, A.M., Nolan, A.J., and Henderson, D.J. (2003). A spatial model with ordinal responses for grazing impact data. *Statistical Modelling*. Submitted.

Lawson, A.B. and Clark, A. (2002). Spatial mixture relative risk models applied to disease mapping. *Statistics in Medicine*, **21**, 359–370.

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

Stolte, A.M., Nolan, A.J., Brewer, M.J., Duff, E.I., Elston, D.A., and Henderson, D.J. (2003). Rapid assessment of grazing and trampling impacts over large areas of rangeland: A sampling approach. *Biological Conservation*. Submitted.

On Stationary GARCH(p,q) Mean Square Stability

Viktorija Carkova[1]

[1] Mathematical Analysis Dpt., University of Latvia, Zeļļuiela 8, LV-1050 Riga, Latvia.

Abstract: The paper deals with symmetric GARCH(p,q) models. Assuming that there exists defined by this model stationary time series we have proposed the necessary and sufficient condition for exponential mean square convergence of any satisfying this model stochastic recurrent procedure to the above stationary time series. A mathematical background of proposal methods are based on the derived by author covariance method for mean square exponential stability analysis of linear stochastic difference equations. That permits to write out a mean square convergence criterion for GARCH(p,q) models with any integer positive p and q in a convenient for application integral form involving the model parameters.

Keywords: GARCH(p,q) models; Stationary Time Series; Stochastic Difference Equations; Mean Square Stability.

1 Introduction: Stationary GARCH(p,q) Models

Over the last decade, there has been a tendency to employ to analysis the financial time-series data model the symmetric regression model for conditional mean, defined by formula

$$Y_t = b_0 + \sum_{k=1}^{n} b_k X_t^{(k)} + \xi_t,\ E\{\xi_t/\Phi_{t-1}\} \equiv 0,\ E\{\xi_t^2/\Phi_{t-1}\} = \sigma_t^2, \qquad (1)$$

with errors (shocks) ξ_t given as GARCH(p,q) model (*Generalized Auto Regressive Conditional Heteroskedasticity*), that takes the following form (Hamilton (1994)):

$$\sigma_t^2 = \theta_0 + \sum_{k=1}^{p} \varphi_k \sigma_{t-k}^2 + \sum_{k=1}^{q} \theta_k \sigma_{t-k}^2 \varepsilon_{t-k}^2. \qquad (2)$$

This process is described for time moments $t \in Z$ by $q+1$ coefficients $\theta_k \geq 0$, $k = 0, 1, ..., q$, p coefficients $\varphi_k \geq 0$, $k = 1, 2, ..., p$, mean b_0, n linear regression coefficients b_k, $k = 1, 2, ..., n$, endogenous and exogenous variables Y_t and $X_t^{(k)}, k = 1, 2, ..., n$ respectively, conditional variance σ_t^2, white-noise type time series $\{\varepsilon_t,\ t \in Z\}$ (that is, i.i.d. random variables with

mean zero and variance one) and the sigma-algebra Φ_{t-1} of information up to time $t-1$, defined by $\{\varepsilon_s,\ s \le t-1\}$. As it has been shown by (Bollerslev (1986)) under assumption

$$\sum_{k=1}^{p} \varphi_k + \sum_{k=1}^{q} \theta_k < 1. \tag{3}$$

there exists defined by (2) stationary time-series $\{\hat{\sigma}_t^2,\ t \in Z\}$ and expectation of deviations $u_t := \sigma_t^2 - \hat{\sigma}_t^2$ of any other satisfying (2) time series $\{\sigma_t^2,\ t \in Z\}$ converge to zero in the mean with $t \to \infty$, that is $\lim\limits_{t\to\infty} E|\sigma_t^2 - \hat{\sigma}_t^2| = 0$. It should be mentioned that parameters of regression model (2) are mainly defined by the least square method and therefore it is preferable (He and Terasvirta (1999)) to analyze a behaviour of the second moments of iterations (2), that is, an asymptotic of sequence $\{E|\hat{\sigma}_t^2 - \sigma_t^2|^2\}$ with $t \to \infty$. We will say that the stationary GARCH model (2) is *exponential mean square stable* if the above second moments exponentially tend to zero as $t \to \infty$, that is, there exist such positive numbers M, λ that

$$\mathbf{E}\big\{|\sigma_t^2 - \hat{\sigma}_t^2|^2\big|_{\sigma_s=\sigma}\big\} \le Me^{-\lambda(t-s)}\mathbf{E}\big\{|\sigma_s^2 - \hat{\sigma}_s^2|^2\big\} \tag{4}$$

for any $t \ge s$, $s \in Z$. The problem arises: to determine a largest set of parameters of model (2), which guarantees the stability property (4). For GARCH(p,1) models this problem has been discussed in our previous paper (Carkova and Gutmanis (2002)). Applying some of well known mathematical results for positive defined matrices, the mentioned paper derives the necessary and sufficient condition for exponential mean square stability in a form of inequality involving forth moment of ε_t and parameters $\varphi_1, ..., \varphi_p, \theta_1$. In spite of the convenience for application of the proposal there approach for $q = 1$, that has been written as an inequality for two specially constructed determinants, it becomes very complicated for GARCH(p,q)-models with $q \ge 2$. To eliminate this lack we will apply another method, developed in paper (Carkova and Carkovs (1969)) for asymptotical stability analysis of linear stochastic difference equations. It permits us under assumption (3) to derive necessary and sufficient exponential mean square stability condition in a form of inequality $E|\varepsilon|^4 < g(\varphi_1, ..., \varphi_p; \theta_1, ..., \theta_q)$.

2 Integral Criteria for GARCH(p,q) Exponential Mean Square Stability

It is easy to write for the deviations $u_t := \hat{\sigma}_t^2 - \sigma_t^2$ the homogeneous difference equation

$$u_t = \sum_{k=1}^{m} a_k u_{t-k} + \sum_{k=1}^{q} \theta_k u_{t-k} y_{t-k}, \tag{5}$$

where $m = \max\{p, q\}$,

$$a_k = \begin{cases} \varphi_k + \theta_k, & \text{if } k \le \min\{p, q\}, \\ \varphi_k, & \text{if } p < k \le q, \\ \theta_k, & \text{if } q < k \le p, \end{cases} \qquad (a_k)$$

and $y_t = \varepsilon_t^2 - 1$. The latter random variables $\{y_t,\, t \in Z\}$ are i.i.d. with mean zero and variance $s^4 := E|\varepsilon_t^2 - 1|^2$ defined by distribution of ε_t. Formula (5) defines a linear difference equation with random coefficients and the problem is: to find necessary and sufficient conditions for exponential mean square decreasing of its solutions. Let sequence $\{u_t,\, t \in Z\}$ be a solution of (5). According to proposal in (V.Carkova and J.Carkovs, 1969) method first of all we have to define two sequences: $\{h_t,\, t \in Z\}$, satisfying for $t > 0$ homogeneous difference equation

$$h_t = a_1 h_{t-1} + a_2 h_{t-2} + \ldots + a_m h_{t-m}, \qquad (h)$$

under conditions $h_0 = 1,\ h_t = 0$ for $t \le -1$, and $\{\tilde{x}_t,\, t > 0\}$ satisfying the same homogeneous difference equation $\tilde{x}_t = a_1\tilde{x}_{t-1} + a_2\tilde{x}_{t-2} + \ldots + a_m\tilde{x}_{t-m}$, but for $t \le 0$ is the same as u_t, that is, $\tilde{x}_t = u_t,\ t \le 0$. It may be proved that owing assumption (3) any of solutions of the above homogeneous equations exponentially tends to zero as $t \to \infty$. Now we should rewrite equation (5) in a following form

$$u_t = \tilde{x}_t + \sum_{s=1}^{t}\sum_{j=1}^{q} h_{t-s}\theta_j y_{s-j}u_{s-j} = g_t + \sum_{j=1}^{q}\sum_{k=1}^{t} h_{t-k-j}\theta_j y_k u_k,$$

where $g_t = \tilde{x}_t + \sum\limits_{j=1}^{q}\sum\limits_{s=1}^{j} h_{t-s}\theta_j y_{s-j}u_{s-j}$ is Φ_0-adopted random sequence for any $t \ge 0$. Squaring the both parts of the above equity and taking a conditional expectation under condition Φ_0 we can reach for conditional second moment $m_t := E\{|u_t|^2/\Phi_0\}$ a following equation

$$m_t = g_t^2 + s^4 \sum_{k=1}^{t} b_{t-k}^2 m_k, \qquad (6)$$

where $b_t = \sum\limits_{j=1}^{q} h_{t-j}\theta_j$. Because g_t^2 and b_t^2 are exponentially decreasing to zero nonnegative sequences any satisfying (6) positive sequence $\{m_t,\, t \ge 0\}$ may be majorized by sequence $\{c^t,\, t \ge 0\}$ for sufficiently large c. Therefore to analyze an asymptotic of this sequence we may apply discrete Laplace transformation multiplying the both parts of (6) by z^t with some constant $z \in (0,\, c^{-1})$ and summarizing by t from 0 to ∞. This approach permits to find function $M(z) := \sum\limits_{t=0}^{\infty} z^t m_t$ in a form of fraction

$$M(z) = \frac{G(z)}{1 - s^4 B(z)}, \qquad (7)$$

where $G(z) := \sum_{t=0}^{\infty} z^t g_t$, $B(z) := \sum_{t=0}^{\infty} z^t b_t^2$. It is obviously that m_t exponentially decreases with $t \to \infty$ if and only if the series $\sum_{t=0}^{\infty} m_t$ converges. Analyzing equality (7) one can make sure of equivalence the latter assertion to inequality $s^4 < (B(1))^{-1}$ involving fourth moment of white noise $s^4 = E\{|\varepsilon_t|^2 - 1|^2\} = E\{|\varepsilon_t|^4\} - 1$ and parameters $\{\varphi_k, \theta_j; k = 1, ..., p, j = 1, ..., q\}$ of model GARPCH(p, q) defined by series $B(1) = \sum_{t=0}^{\infty} b_t^2$. Let $B_1(z)$ be a discrete Laplace transformation of sequence $\{b_t\}$, that is, $B_1(z) := \sum_{t=0}^{\infty} b_t z^t$. Applying well known Cauchy theorem one may calculate any term of sequence $\{b_t\}$ as contour integral of complex valued function $B_1(z)$ multiplied by z^{-t-1}:

$$b_t = \frac{1}{2\pi i} \int\limits_{|z|=1+\delta} B(z) z^{-1-t} dz,$$

with an arbitrary sufficiently small by module real number δ. Therefore formula $B(1) = \sum_{t=0}^{\infty} b_t^2$ may written in an integral form:

$$B(1) = \sum_{t=0}^{\infty} \left(\frac{1}{2\pi i} \int\limits_{|z|=1+\delta} B_1(z) z^{-1-t} dz \right) \left(\frac{1}{2\pi i} \int\limits_{|x|=1} B_1(x) x^{-1-t} dx \right) =$$

$$= -\frac{1}{4\pi^2} \int\limits_{|z|=1+\delta} \int\limits_{|x|=1} B_1(z) B_1(x) \frac{dxdz}{xz-1} = \frac{1}{2\pi i} \int\limits_{|z|=1} B_1(z) B_1(z^{-1}) z^{-1} dz.$$

The function $B_1(z)$ is a Z-transformation of series $b_t = \sum_{j=1}^{q} h_{t-j}\theta_j$ formed by solution $\{h_t\}$ of difference equation (h). Let $H(z)$ be a Z-transformation of this solution. Owing to special chosen initial conditions for that, the delayed solution $\{h_{t-k}\}$ with any $k \geq 0$ has Z-transformation $z^k H(z)$. Therefore applying Z-transformation to formula (h) one can find $B_1(z) = \left(\sum_{j=1}^{q} \theta_j z^j \right) \Big/ \left(1 - \sum_{k=1}^{m} a_k z^k \right)$. Finally defined by (9) expression $B(1)$ has an integral form

$$B(1) = \frac{1}{2\pi i} \int\limits_{|z|=1} \frac{(\sum_{j=1}^{q} \theta_j z^j)(\sum_{j=1}^{q} \theta_j z^{q-j})}{z(1 - \sum_{k=1}^{m} a_k z^k)(1 - \sum_{k=1}^{m} a_k z^{m-k})} z^{m-q-1} dz, \qquad (8)$$

where $m = \max\{p, q\}$ and a_k defined above in formula (a_k). Therefore the necessary and sufficient condition for stationary GARCH(p,q) mean

square stability has a form an inequality $s^4 < B(1)^{-1}$ with $B(1)$ defined by formula (8). Because under assumption (3) the absolute values of all roots of polynomial $\sum_{k=1}^{m} a_k z^{m-k}$ are less than 1, the integral in (10) can be calculated applying residual theory.

Example. Let we should deal with GARCH(2,2) model. Then $p = q = 2$, $H(z) = (1 - (\varphi_1 + \theta_1)z - (\varphi_2 + \theta_2)z^2)^{-2}$, and integral (9) is equal to

$$\frac{1}{2\pi i} \int\limits_{|z|=1} \frac{z(\theta_1^2 + \theta_2^2) + (1 + z^2)\theta_1\theta_2}{(1 - (\varphi_1 + \theta_1)z - (\varphi_2 + \theta_2)z^2)(z^2 - (\varphi_1 + \theta_1)z - (\varphi_2 + \theta_2))} dz.$$

Not so difficult to find roots of polynomial $z^2 - (\varphi_1 + \theta_1)z - (\varphi_2 + \theta_2)$ (which should be less than one by module) and applying residual theory for the above integral to write necessary and sufficient exponential mean square stability condition of GARCH(2,2) in a following complete form:

$$s^4 < \frac{2[(1 - \varphi_2 - \theta_2)^2 - (\varphi_1 + \theta_1)^2](1 + \varphi_2 + \theta_2)}{(\theta_1^2 + \theta_2^2)(1 - \varphi_2 - \theta_2) + 4(\varphi_1 + \theta_1)\theta_1\theta_2}.$$

References

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics,* **31**, 307–327.

Carkova, V. and Gutmanis, N. (2002). On Convergence of GARCH(p,q). In: *Statistical Modelling in Society. Proceedings of the 17th International Workshop on Statistical Modelling (Chania, Greece, 8-12 July 2002)*. National and Kapodistrian University of Athens & University of North London, 149–152.

Carkova, V. and Carkovs, J. (1969). On Stability of Solutions of Difference Equations with Random Coefficients. *Latvijskij Matematicheskij Ezhyegodnik*, **5**, 153–173. Riga: LU.

Hamilton, J.D. (1994). *Time Series Analysis*. Princeton: Princeton University Press.

He, C. and Terasvirta, T. (1999). Fourth moment structure of the GARCH (p; q) process. *Econometric Theory,* **15**, 824–846.

Ling, S. (1999). On the stationarity and the existence of moments of conditional heteroskedastic ARMA models. *Statistica Sinica,* **9**, 1119–1130.

On Price Stochastic Equilibrium

Jevgenijs Carkovs[1] and Remigijs Počs[2]

[1] Information Technology Institute, Riga Technical University, Meža iela 1/4, Riga, LV-1048, Latvia
[2] International Business & Custom Institute, Riga Technical University, Indriķu 8, Riga, LV-1004, Latvia

Abstract: The paper proposes a stochastic analysis approach to equilibrium stability analysis of an adaptive Samuel-Marshall type single component market. Assuming that an equilibrium can be achieved by the equality of demand to supply, manufacturer would like to stabilize the price of a product unit into a small neighborhood of the level $\bar{p}$. Because to enter the market a manufacturer needs a time, he can manage by a chosen supplied quantity at the delayed time moment $t - \tau(t)$. Taking into account this circumstance and permanent random perturbations of demand elasticity function the paper discusses stochastic price dynamics. Our approach is based on asymptotical theorems of stochastic calculus and second Lyapunov methods for stability analysis of stochastic functional differential equations. That allows not only to estimate how a time delay of supply and other market performances exert price dynamics, but also to advance in stochastic stability analysis, calculating dependence of maximal admissible price equilibrium volatility on demand and supply elasticity functions.

Keywords: Price equilibrium; Stochastic stability; Adaptive market.

1 Introduction: Adaptive Single Component Market

The paper deals with simplest mathematical model of an adaptive Samuel-Marshall type single component market under assumption that a manufacturer has a monopoly there and he would like to stabilize the price of a product unit into a small neighborhood of the level $\bar{p}$. Let us remind, that in any classical single component market model supply S_t and demand D_t are dependent on a price of product unit history up to time $\mathcal{F}^t := \{p(s),\, s \leq t\}$ and equilibrium $p(t) \equiv \bar{p}$ may be achieved by the equality of demand $D(\bar{p})$ to supply $S(\bar{p})$ (see, for example, Hamilton (1994), Samuelson (1974), Sharpe (1964)). As in the classical Samuelson model we will suppose equilibrium to be reached due to an adaptive price dynamical property: the price movement $(\Delta p)(t) := p(t + \triangle) - p(t)$ is proportional to difference $D_t - S_t$ multiplied by time increment $\triangle$. To control a price $p(t)$ a manufacturer should keep a supplied quantity S_t at the level of demand quantity D_t, but to enter the market at the time moment t he needs some time $\tau(t)$. Therefore manufacturer has a delayed reaction because he

is guided by the price at the moment of time $t-\tau(t)$. As a result the supply S_t is a function of a delayed price $p(t-\tau(t))$, that is $S_t = S(p(t-\tau(t)))$. Because supply quantity is formed by manufacturer based on known price value p we may suppose that supply is deterministic sufficiently smooth function and model this function within a small equilibrium vicinity in a linear form

$$S(p) = bp + \beta. \tag{1}$$

But demand D_t at time moment t does not depend on manufacturer. He can find this function analyzing statistical data $\{D_s, p(s),\ s \le t\}$ only and applying some of well derived regression procedures. Let us suppose for simplicity that demand D_t has a quick response on price dynamics and is dependent only on present price value $p(t)$. Therefore for a given price history $\mathcal{F}^{\sqcup}$ we may model a deterministic part of demand as a conditional expectation $D(p(t)) := \mathcal{E}\{\mathbf{D_t}/\mathcal{F}^{\sqcup}\}$ and forecast mean value of a demand quantity for a given price value p as a function $D(p)$. The above assumptions permit to analyze price dynamics writing out market mathematical model in a form of the first order ordinary differential equation with delay

$$\frac{dp(t)}{dt} = D(p(t)) - S(p(t-\tau(t)). \tag{2}$$

We should take into consideration that Samuelson's adaptive assertion about adaptive price increments dynamics is only *rational expectation* of market price reaction on supply and demand values and therefore equation (2) reflects price dynamics in the mean. Unfortunately real price mostly is very irregular function of time t. Recent decades has appears many papers which intensively developed the branch of modern economics concerning the price dynamics analysis and elaboration of a rational algorithm of investor behavior, taking into account the financial market statistical uncertainty. It has been shown that it is not enough to know smooth dynamical performances of financial flows, reached by moving-average procedure, but also is necessary to analyze extremely complicated and bad predictable chaotic price oscillations. This made many researchers use Ito stochastic calculus for modeling price dynamics. As an example one can specify the well-known Black-Scholes option-pricing formula used not only by scientists in the theoretical financial economics but also by most of brokers for gambling on a stock exchange (see, for example, Black and Scholes (1973), Marshall (1979), Sharpe (1964)). This paper also deals with the stochastic analysis of price dynamics, writing an adaptive Samuelson's assertion in a form of stochastic Ito equation

$$dp(t) = (D(p(t)) - S(p(t-\tau(t)))dt + \sigma p(t) dB(t), \tag{3}$$

where $B(t)$ is standard Brawn motion process, and parameter σ (called by *volatility*) allows to take into account value of risk connected with this model of price dynamics (see, for example, Sharpe (1964)).

2 Stochastic Analysis Approach

Let us suppose that price equilibrium can be reached owing to equality of demand and supply in the mean, that is $\mathbf{E}\{dp/\mathcal{F}^{\sqcup}\} = \prime$. First of all we will discuss the dependence of equilibrium stability on demand elasticity by price (parameter b), fraction of demand elasticity by price and supply elasticity by price (parameter $c = a/b$ with $a > 0$, $b < 0$, $-b < a$), volatility $\sigma \geq 0$, and time-delay $\tau(t)$ probabilistic properties, writing for price deviations $x(t) := p(t) - \bar{p}$ linear stochastic Ito equation in a following form

$$dx(t) = b(cx(t) - x(t - \tau(t)))dt + \sigma x(t)dw(t). \tag{4}$$

Under condition of independence of the involving in the above equation processes $B(t)$, $\tau(t)$ and strong mixing condition for stationary process $\tau(t)$ our paper proves that for equilibrium stability it is necessary a negativity of real parts of all roots of function $v(z) := z - b(c - M(-z))$, where $M(z)$ is moment function of time delay. For example, if stationary distribution of time delay is given by formulae $\mathbf{P}\{\tau(t) = 0\} = \hat{\pi}$, $\mathbf{P}\{\tau(t) = 1\} = 1 - \hat{\pi}$, the necessary price equilibrium stability condition has a following form (Šadurskis and Tsarkov (2001)):

$$\hat{\pi} < \frac{1-c}{2}, \; b < \frac{\arccos((c-\hat{\pi})/(1-\hat{\pi}))}{\sqrt{(1-c)(1+c-2\hat{\pi})}}. \tag{5}$$

It should be mentioned that this inequality guarantees exponential decreasing only for price mean value, but price variance may be infinitely increasing. Applying proposal in Šadurskis and Tsarkov (2001) asymptotic methods and covariance approach for mean square stability analysis of stochastic functional differential equations (see, for example, Tsarkov (1989)) one can find necessary and sufficient condition for decreasing of price variance to zero with $t \to \infty$ in a form of inequality for volatility

$$\sigma^2 < \pi\Big(\int_0^{\infty} \frac{dz}{|iz - b(c - \chi(-z))|^2}\Big)^{-1}, \tag{6}$$

where $\chi(z)$ is characteristic function of stationary random delay $\tau(t)$. For above mentioned example of binomial market this inequality has a form (Carkovs and Počs (2002)):

$$\sigma^2 < \sigma_{cr}^2 := \frac{2(1-\hat{\pi})b\nu(\nu - \sin(b\nu))}{(1-\hat{\pi})\cos(b\nu) + c - \hat{\pi}} \tag{7}$$

where $\nu = \sqrt{(1-c)(1-2\hat{\pi}+c)}$.

3 Price Nonlinear Dynamics

Linear differential market model even in a more complicated nonhomogeneous form than (4)

$$dp(t) = \{ap(t) - bp(t - \tau(t)) + \alpha - \beta + \eta(t)\}dt + \sigma P(t)dB(t), \qquad (8)$$

where $\eta(t)$ is zero-mean stationary process, satisfactory fits price dynamics only for parameters located within stability region and sufficiently far from border of stability. In this case one can forecast price dynamics as a steady-state stationary process with mean $\bar{p} = (\alpha - \beta)/(b - a)$ and spectral density which is proportional to spectral density $f_\eta(\lambda)$ of noise $\eta(t)$ and inverse proportional to squared modules of frequency response

$$V(\lambda) = |\lambda - b\mathrm{Im}\{M(i\lambda)\}|^2 + |a - b\mathrm{Re}\{M(i\lambda)\}|^2,$$

where $M(z)$ is moment generating function of stationary delay $\tau(t)$. But for parameters a and b lying on the border of stability function $V(\lambda)$ becomes equal to zero for some values of λ and the steady-state stationary of equation (8) does not exist. Therefore chosen linear model becomes invalid because predictable by this model price dynamics has such a large scatter that $p(t)$ leaves a chosen vicinity of equilibrium $\bar{p}$. That is why for the parameters near border of stability one should pass to more complicated nonlinear mathematical model. Let for example market model has delayed linear supply (1) with stationary distributed binomial time delay $\mathbf{P}\{\tau(t) = 1\} = 1 - \hat{\pi}$, $\mathbf{P}\{\tau(t) = 0\} = \hat{\pi}$ and nonlinear demand $D_t = D(p(t)) + \eta(t)$. This model may be written in a form of an ordinary functional differential equation

$$\frac{dp(t)}{dt} = D(p(t)) - bp(t - \tau(t)) - \beta + \eta(t),$$

where $\eta(t)$ is stationary process with known correlation function. As it was mentioned above the parameters $(b, D'(\bar{p}) := a, \hat{\pi})$ should be chosen near a critical point $\{b_{cr}, a_{cr}, \pi_{cr}\}$ laying on the border of stability. Taking $a = a_{cr}$, $\hat{\pi} = \pi_{cr}$, and $b = b_{cr} + \varepsilon$, where ε is a small parameter, and applying a stochastic averaging procedure (see, for example, Šadurskis and Tsarkov (2001)) under condition $D'(\bar{p}) > 0$, $D''(\bar{p}) < 0$ one can predict asymptotically stable price stochastic oscillations

$$p(t) = \bar{p} + (\bar{r} + \sqrt{\varepsilon}\rho(\varepsilon t))\cos(\nu t + \varphi(\varepsilon t)), \qquad (9)$$

with frequency $\nu = \sqrt{(a - b)(a + b - 2a\pi_{cr})}$, where $\rho(t)$ is stationary Gaussian process of Ornstain-Ulenbeck type and $\varphi(t)$ is stationary Gaussian process given on the circle $\mathcal{S}^1$ by stochastic Ito equation with zero drift and constant diffusion. Corresponding to price dynamics (9) steady-state process approximately may be represented as a limit cycle on the Demand-Supply phase plane called by stable stationary price business cycle.

References

Black, F. and Scholes, M. (1973). The pricing of options and vorporate liabilities. *Journal of Political Economy,* **81**, 637–657.

Carkovs, J. and Počs, R. (2002). Price equilibrium volatility reserve for Marshal-Samuelson adaptive market. In: *Statistical Modelling in Society. Proceedings of the 17th IWSM (Chania, Greece, 8-12 July 2002).* 153–156.

Marshall, A. (1979). *Principles of Economics.* New York: MacMillan Press Ltd.

Merton, R. (1990). *Continuous Time Finance.* Cambridge, UK: Blackwell.

Šadurskis, K. and Tsarkov, Y. (2001) Asymptotic methods for stability analysis of retarded dynamical systems with Markov switching. In: *Proceedings of ASMDA'2001, v.2/2.* 905–910.

Samuelson, P. (1974). *Foundations of Economic Analysis.* New York: Atheneum.

Sharpe, W. (1964). Capital asset prices: a theory of market equilibrium under conditions of risk. *Journal of Finance*, **19**,425–442.

Tsarkov, Y. F. (1989). *Random Perturbations of Functional Differential Equations.* Riga: Zinatne.

On the Estimation of Parameters in the Chemical Balance Weighing Design under the Vovariance Matrix of Errors $\sigma^2\mathbf{G}$

Bronislaw Ceranka[1] and Malgorzata Graczyk[1]

[1] Department of Mathematical and Statistical Methods Agricultural University Wojska Polskiego 28, 60-637 Poznań, Poland Email: bronicer@owl.au.poznan.pl Email: magra@owl.au.poznan.pl

Abstract: The paper is studying the estimation problem of individual weights of objects using a chemical balance weighing design under the restriction on the number times in which each object is weighed. We assume that errors have the same variances and they are correlated. The necessary and sufficient conditions under which the lower bound of variance of each of the estimated weights is attained are given.

Keywords: Chemical balance weighing design.

1 Introduction

Let us consider the class $\mathbf{\Phi}_{n\times p,m}(-1,0,1)$ of the $n \times p$ matrices $\mathbf{X}$ with elements equal to $-1, 0$ or 1, where m there is the maximum number of elements equal to -1 and 1 in each column of the matrix $\mathbf{X}$. The matrices belonging to this class there are the design matrices of the chemical balance weighing designs. Suitable model we can write in the form,

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e}, \tag{1}$$

where $\mathbf{y}$ is an $n \times 1$ random observed vector of the recorded results of weights, $\mathbf{w}$ is an $p \times 1$ column vector representing unknown weights of objects and $\mathbf{e}$ is an $n\times 1$ random vector of errors. We assume that $\mathrm{E}(\mathbf{e}) = \mathbf{0}_n$ and the errors are correlated and with equal variances, i.e. $\mathrm{Var}(\mathbf{e}) = \sigma^2\mathbf{G}$, where $\mathbf{0}_n$ is the $n \times 1$ column vector of zeros, $\mathbf{G} = g\left[(1-\rho)\mathbf{I}_n + \rho\mathbf{1}_n\mathbf{1}_n'\right]$, where $g > 0$, $-1 < \rho < 1$ are given constants.
For estimating individual unknown weights of objects we can use the normal equations

$$\mathbf{X}'\mathbf{G}^{-1}\mathbf{X}\hat{\mathbf{w}} = \mathbf{X}'\mathbf{G}^{-1}\mathbf{y}, \tag{2}$$

where $\hat{\mathbf{w}}$ is the vector of the weights estimated by the least squares method. The chemical balance weighing design is singular or nonsingular depending on whether the matrix $\mathbf{X}'\mathbf{G}^{-1}\mathbf{X}$ is singular or nonsingular, respectively. If

$\mathbf{X}'\mathbf{G}^{-1}\mathbf{X}$ is nonsingular the least squares estimator of $\mathbf{w}$ is given in the form,

$$\hat{\mathbf{w}} = (\mathbf{X}'\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{G}^{-1}\mathbf{y} \tag{3}$$

and the variance - covariance matrix of $\hat{\mathbf{w}}$ is given by formula,

$$\mathrm{Var}(\hat{\mathbf{w}}) = \sigma^2(\mathbf{X}'\mathbf{G}^{-1}\mathbf{X})^{-1}. \tag{4}$$

In the case $\mathbf{G} = \mathbf{I}_n$, Hotelling (1944) has studied some problems connected with chemical balance weighing designs. He has shown that for the chemical balance weighing design the minimum attainable variance for each of the estimated weights is σ^2/n. He proved the theorem that each of the variance of the estimated weights attains the lover bound if and only if $\mathbf{X}'\mathbf{X} = n\mathbf{I}_p$. This design is called the optimum chemical balance weighing design. It implies that for the optimum chemical balance weighing design the elements of the matrix $\mathbf{X}$ are -1 and 1, only. In this case, several methods of construction the optimum chemical balance weighing designs are available in Raghavarao (1971) and Banerjee (1975). In the model of optimum chemical balance weighing design with equal correlated errors Ceranka and Katulska (1998) gave the sufficient and necessary conditions under which the lower bound of variances of the estimators was attained.

2 Variance Limit of Estimated Weights

Let assume that the design matrix $\mathbf{X} \in \mathbf{\Phi}_{n\times p,m}(-1,0,1)$ is of full column rank p, $\mathbf{c}$ denote an $p \times 1$ vector. From Section 1c.1 (ii) (b) in Rao (1973) we get:

Lemma 2.1 For the design matrix $\mathbf{\Phi}_{n\times p,m}(-1,0,1)$ of rank p, any symmetric positive definite $n \times n$ matrix $\mathbf{G}$ and $p \times 1$ vector $\mathbf{c}$ inequality

$$\mathbf{c}'(\mathbf{X}'\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{c} \geq \frac{(\mathbf{c}'\mathbf{c})^2}{\mathbf{c}'(\mathbf{X}'\mathbf{G}^{-1}\mathbf{X})\mathbf{c}}$$

is true and the equality is fulfilled if and only if $\mathbf{c}$ there is eigenvector of the matrix $\mathbf{X}'\mathbf{G}^{-1}\mathbf{X}$.

Lemma 2.2 The matrix $(1-\rho)\mathbf{I}_n + \rho\mathbf{1}_n\mathbf{1}_n'$ is positive definite if and only if $\frac{-1}{n-1} < \rho < 1$.

It is obvious that if $\frac{-1}{n-1} < \rho < 1$ and $g > 0$ then the matrix $\mathbf{G} = g[(1-\rho)\mathbf{I}_n + \rho\mathbf{1}_n\mathbf{1}_n']$ is positive definite and the matrix $\mathbf{X}'\mathbf{G}^{-1}\mathbf{X}$ is nonsingular if and only if the matrix $\mathbf{X}'\mathbf{X}$ is nonsingular, i.e. if and only if $\mathbf{X}$ is of full column rank $(= p)$.

Let assume that the matrix $\mathbf{G}$ is given as

$$\mathbf{G} = g[(1-\rho)\mathbf{I}_n + \rho\mathbf{1}_n\mathbf{1}_n'], \tag{5}$$

where $\frac{-1}{n-1} < \rho < 1$ and $g > 0$.

Theorem 2.1 In the 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 by (5), the variance of $\hat{w}_j$ for a particular j, such that $j = 1,2,...,p$, cannot be less then

$$\mathrm{Var}(\hat{w}_j) \geq \begin{cases} \frac{\sigma^2 g(1-\rho)}{m} & \text{if} \quad 0 \leq \rho < 1, \\ \frac{\sigma^2 g(1-\rho)}{m-\frac{\rho}{1+\rho(n-1)}(m-2u)^2} & \text{if} \quad \frac{-1}{n-1} < \rho < 0, \end{cases} \qquad j = 1,2,...,p$$

where $m = \max\{m_1, m_2, ..., m_p\}, m_j$ represents the number of elements equal to -1 and 1 in jth column of $\mathbf{X}$, $u = \min\{u_1, u_2, ..., u_p\}, u_j$ represents the number of elements equal to -1 in jth column of $\mathbf{X}$, $j = 1,2,...,p$.

Proof: Let $\mathbf{c}_j$, $j = 1,2,...,p$, be the vector equal to jth column of the matrix $\mathbf{I}_p$. Then we have
$\hat{w}_j = \mathbf{c}_j'\hat{\mathbf{w}}$ and $\mathrm{Var}(\hat{w}_j) = \sigma^2\mathbf{c}_j'(\mathbf{X}'\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{c}_j, \qquad j = 1,2,...,p.$
Since the matrix $\mathbf{X} \in \mathbf{\Phi}_{n\times p,m}(-1,0,1)$ is of full column rank and

$$\mathbf{G}^{-1} = \frac{1}{g(1-\rho)}[\mathbf{I}_n - \frac{\rho}{1+\rho(n-1)}\mathbf{1}_n\mathbf{1}_n'],$$

then from the lemmas 2.1 and 2.2 we have

$$\mathrm{Var}(\hat{w}_j) \geq \sigma^2\frac{(\mathbf{c}_j'\mathbf{c}_j)^2}{\mathbf{c}_j'\mathbf{X}'\mathbf{G}^{-1}\mathbf{X}\mathbf{c}_j} =$$

$$= \sigma^2 g(1-\rho)\frac{1}{\mathbf{c}_j'\mathbf{X}'\mathbf{X}\mathbf{c}_j - \frac{\rho}{1+\rho(n-1)}\mathbf{c}_j'\mathbf{X}'\mathbf{1}_n\mathbf{1}_n'\mathbf{X}\mathbf{c}_j} \geq$$

$$\geq \sigma^2 g(1-\rho)\frac{1}{m - \frac{\rho}{1+\rho(n-1)}\mathbf{c}_j'\mathbf{X}'\mathbf{1}_n\mathbf{1}_n'\mathbf{X}\mathbf{c}_j} \geq$$

$$\geq \begin{cases} \frac{\sigma^2 g(1-\rho)}{m} & \text{if} \quad 0 \leq \rho < 1, \\ \frac{\sigma^2 g(1-\rho)}{m-\frac{\rho}{1+\rho(n-1)}(m-2u)^2} & \text{if} \quad \frac{-1}{n-1} < \rho < 0, \end{cases} \qquad j = 1,2,...,p. \qquad (6)$$

Since elements $x_{ij} = -1, 1$ or 0 only, hence the thesis.
In the special case $m = n$ and $\mathbf{X} \in \mathbf{\Psi}_{n\times p}(-1,1)$, the class of the $n \times p$ matrices $\mathbf{X}$ with elements equal to -1 and 1, theorem 2.1 was given in Ceranka and Katulska (1998). When $m = n, \rho = 0, g = 1$ and $\mathbf{X} \in \mathbf{\Psi}_{n\times p}(-1,1)$ the theorem 2.1 was proved in Hotelling (1944).

Definition 2.1 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 the matrix $\mathbf{G}$ is of the form (5), is said to be

optimal for the estimated individual weights if the variance of each of the estimators attains the lower bound, i.e. if

$$\mathrm{Var}(\hat{w}_j) = \begin{cases} \frac{\sigma^2 g(1-\rho)}{m} & \text{if} \quad 0 \le \rho < 1, \\ \frac{\sigma^2 g(1-\rho)}{m - \frac{\rho}{1+\rho(n-1)}(m-2u)^2} & \text{if} \quad \frac{-1}{n-1} \le \rho < 0, \end{cases} \qquad j = 1, 2, ..., p.$$

Now, we give the necessary and sufficient conditions under which the lower bound is attained.

Theorem 2.2 Let $0 \le \rho < 1$. 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 the matrix $\mathbf{G}$ is of the form (5), is optimal for the estimated individual weights if and only if
(i) $\mathbf{X}'\mathbf{X} = m\mathbf{I}_p$ and
(ii) $\mathbf{X}'\mathbf{1}_n = \mathbf{0}_p$.

Proof: To prove the necessity part let notice that from lemma 2.2 the first inequality in (6) is equality if and only if $\left[\mathbf{X}'\mathbf{X} - \frac{\rho}{1+\rho(n-1)}\mathbf{X}'\mathbf{1}_n\mathbf{1}_n'\mathbf{X}\right]\mathbf{c}_j = \mu_j\mathbf{c}_j,\ \mu_j > 0,\ j = 1, 2, ..., p$. The second inequality in (6) is equality for each j if and only if $\mathbf{c}_j'\mathbf{X}'\mathbf{X}\mathbf{c}_j = m$ for $j = 1, 2, ..., p$. The third inequality is equality if and only if $\mathbf{c}_j'\mathbf{X}'\mathbf{1}_n = 0$ for $j = 1, 2, ..., p$. These conditions imply that $\mathbf{X}'\mathbf{X} = diag\{\mu_1, \mu_2, ..., \mu_p\}, \mu_1 = \mu_2 = ... = \mu_p = m$ and $\mathbf{X}'\mathbf{1}_n = \mathbf{0}_p$. Finally we get (i) and (ii). The proof of sufficiency part is obvious.

Theorem 2.3 Let $\frac{-1}{n-1} < \rho < 0$. 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 the matrix $\mathbf{G}$ is of the form (5), is optimal for the estimated individual weights if and only if
(i) $\mathbf{X}'\mathbf{X} = m\mathbf{I}_p - \frac{\rho(m-2u)^2}{1+\rho(n-1)}(\mathbf{I}_p - \mathbf{1}_p\mathbf{1}_p')$,
(ii) $u_1 = u_2 = ...u_p = u$ and
(iii) $\mathbf{X}'\mathbf{1}_n = \mathbf{z}_p$,
where $\mathbf{z}_p$ is $p \times 1$ vector, for which the jth element is to equal $(m - 2u)$ or $-(m - 2u), j = 1, 2, ..., p$.

Proof: The proof of the necessity part is similarly to the proof of two first inequalities in the theorem 2.2. The third inequality in (6) is equality if and only if $\mathbf{c}_j'\mathbf{X}'\mathbf{1}_n = (m - 2u)$ or $-(m - 2u)$ for $j = 1, 2, ..., p$. It implies $\mathbf{X}'\mathbf{X} - \frac{\rho}{1+\rho(n-1)}\mathbf{X}'\mathbf{1}_n\mathbf{1}_n'\mathbf{X} = diag\{\mu_1, \mu_2, ..., \mu_p\}, \mu_1 = \mu_2 = ... = \mu_p = m - \frac{\rho(m-2u)^2}{1+\rho(n-1)}$ and $\mathbf{X}'\mathbf{1}_n = \mathbf{z}_p$.

The proof of sufficiency part is obvious. Hence the thesis.
Let us consider the case $0 \le \rho < 1$. 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 as (5), is optimal if and only if the conditions (i) and (ii) from the theorem 2.2 hold. From this second condition we get

Corollary 2.1 Let $0 \leq \rho < 1$. The necessary condition for existence optimum 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 (5), is $m \equiv 0(mod 2)$.
The condition (i) from the theorem 2.1 is the same as condition determining optimum chemical balance weighing design with the design matrix $\mathbf{X} \in \mathbf{\Phi}_{n\times p,m}(-1,0,1)$ and with variance - covariance matrix of errors $\sigma^2\mathbf{I}_n$. Then we have

Corollary 2.2 Let any $p_1(\leq p)$ columns of the design matrix $\mathbf{X} \in \mathbf{\Phi}_{n\times p,m}(-1,0,1)$ of the optimum chemical balance weighing design with the variance - covariance matrix $\sigma^2\mathbf{I}_n$ form new design matrix $\mathbf{X}^* \in \mathbf{\Phi}_{n\times p,m}(-1,0,1)$ of the chemical balance weighing design. Chemical balance weighing design with the design matrix $\mathbf{X}^* \in \mathbf{\Phi}_{n\times p,m}(-1,0,1)$ with the variance - covariance matrix of errors $\sigma^2\mathbf{G}$, where $\mathbf{G}$ is given by (5) and $0 \leq \rho < 1$, is optimal if and only if $\mathbf{X}'\mathbf{1}_n = \mathbf{0}_p$.

Theorem 2.4 Let $0 \leq \rho < 1$. The existence of the optimum chemical balance weighing design with the 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 by (5), is equivalent to the existence of the optimum chemical balance weighing design with the matrix $\mathbf{X}^* \in \mathbf{\Phi}_{2n\times 2p,m}(-1,0,1)$ in the form $\mathbf{X}^* = \begin{bmatrix} \mathbf{X} & \mathbf{X} \\ \mathbf{X} & -\mathbf{X} \end{bmatrix}$ and with the variance - covariance matrix of errors $\sigma^2\mathbf{G}^*$, where $\mathbf{G}^* = g[(1-\rho)\mathbf{I}_{2n} + \rho\mathbf{1}_{2n}\mathbf{1}'_{2n}]$.

Proof: Let notice that if $\mathbf{X}$ there is the matrix of optimum chemical balance weighing design with the variance - covariance matrix of errors $\sigma^2\mathbf{G}$, where $\mathbf{G}$ is given by (5), then $\mathbf{X}\mathbf{X}' = m\mathbf{I}_p$ and $\mathbf{X}'\mathbf{1}'_n = \mathbf{0}_p$. These conditions are fulfilled if and only if $\begin{bmatrix} \mathbf{X} & \mathbf{X} \\ \mathbf{X} & -\mathbf{X} \end{bmatrix}' \begin{bmatrix} \mathbf{X} & \mathbf{X} \\ \mathbf{X} & -\mathbf{X} \end{bmatrix} = 2m\mathbf{I}_{2p}$ and $\mathbf{X}'\mathbf{1}_n = \mathbf{0}_p$. Hence the thesis.

References

Banerjee, K.S. (1975). *Weighing Designs for Chemistry, Medicine, Economics, Operations Research, Statistics.* New York: Marcel Dekker.

Ceranka, B. and Katulska, K. (1998). *Optimum chemical balance weighing designs under equal correlations of errors.* Moda 5-Advances in Model

Oriented Data Analysis and Experimental Design, A. C. Atkinson, L. Pronzato, H. P. Wynn, Eds., Physica Verlag, Heidelberg, 3–9.

Hotelling, H. (1944). Some improvements in weighing designs and other experimental techniques. *Annals of Mathematical Statistics*, **15**, 297–305.

Raghavarao, D. (1971). *Constructions and Combinatorial Problems in Designs of Experiments.* New York: Wiley.

Rao, C.R. (1973). *Linear Statistical Inference and its Applications, Second Edition.* New York: Wiley.

Informative Drop-out Model for Longitudinal Binary Data using Bayesian Approach

Jennifer S. K. Chan[1] and Vicki K. K. Chau[2]

[1] The University of Hong Kong, Department of Statistics and Actuarial Science, Pokfulam, Hong Kong
[2] The University of Hong Kong, Department of Statistics and Actuarial Science, Pokfulam, Hong Kong

Abstract: A model is proposed for longitudinal binary data with informative drop-out. The model combines a conditional AR1 model for the underlying response with a logistic regression model for the drop-out process. Parameter estimation is done through a Bayesian approach. The model is demonstrated and compared through a methadone clinic data set. It is anticipated that information gathered from modeling the drop out process has practical implications for the interpretation of the data.

Keywords: Informative drop-out; Longitudinal binary data; Bayesian hierarchy; Markov Chain Monte Carlo; Gibbs sampler.

1 Introduction

The collection of longitudinal binary data is common in clinical trials or longitudinal studies when repeated measurements, positive or negative to certain tests, are made on the same subject over time. Since many longitudinal studies are lengthy, subjects undergoing longitudinal studies may drop-out prematurely, resulting in a large class of distinct missingness patterns. One important issue arising from the problem of drop-out is whether the drop-out process is related to the measurement process. Drop-out processes can be classified into three types: completely random, random and informative drop-out (Rubin,1976, Little and Rubin, 1987). Completely random and random drop-out are often referred to as being ignorable which indicates that it is not necessary to specify a model for drop-out in a likelihood-based analysis of the measurement process. Informative drop-out (ID), on the other hand, is said to be non-ignorable as the drop-out mechanism cannot be ignored when estimating parameters for the data. Diggle and Kenward (1994) have demonstrated that there are biases in the parameter estimated if such drop-out mechanism is not accommodated in the model. Special modeling strategies are therefore required for inference when the drop-out process is informative.

2 Modeling Strategy

A modeling strategy for longitudinal binary data with informative drop-out has been proposed. A conditional AR1 model with random intercepts that accounts for population heterogeneity was proposed for the underlying response and a logistic regression model for the drop-out process. Both the probability of positive outcomes and that of drop-out were assumed to be logit linear in some covariates and outcomes.

2.1 Background of Data Set

We will use a methadone clinic data set reported by Chan *et al.* (1998) to demonstrate our models. This data set is a record of drug users enrolled in a methadone maintainance treatment (MMT) programme at a clinic in Western Sydney in 1986. The record consists of several information including drug user's weekly urine test result that are positive ($Y = 1$) or negative ($Y = 0$) to morphine, a biological marker for heroin use, the dosage of methadone d in milligram (mg) at the time of urine test and also their duration of treatment in weeks t. There are 136 heroin users, submitting a total of 2,872 urine screens with 16% of them being positive for heroin. The dosage of methadone averaged over the 2,872 incidents is 64mg. Each user submitted 4 to 26 weekly outcomes and the average number of treatment weeks per heroin user is 21.1 weeks. 51 drug users dropped out before the end of 26 weeks and the rest having 26 outcomes were regarded as having completed the program. For all analyses, each urine screen result rather than each patient served as the unit of analysis.

2.2 Models

Let Y_{it} denote the binary outcome for patient i in week t. The vector of all possible outcomes for patient i can be separated into

$$\boldsymbol{Y}_i = (Y_{it})' = (\underbrace{Y_{i1}, \cdots, Y_{i,n_i}}_{\text{Observed}\boldsymbol{Y}'_{oi}}, \underbrace{Y_{i,n_i+1}, \cdots, Y_{i,n}}_{\text{Unobserved}\boldsymbol{Y}'_{mi}})'$$

where n_i denotes the number of observed Y_{it} and the vector of all outcomes is denoted by $\boldsymbol{Y'} = (\boldsymbol{Y'}_1, \boldsymbol{Y'}_2, \cdots, \boldsymbol{Y'}_I)$.
Similarly, let R_{it} denote the drop-out indicator for patient i in week t such that $R_{it} = I(t > n_i)$ where R_{it}=1 if Y_{it} is unobserved ($t > n_i$) and zero otherwise. Then the vector of all drop-out indicators R for patient i is $\boldsymbol{R}_i = (R_{it})'$ which is a series of n_i '0' followed by 26 - n_i '1'.
For the outcome model, the conditional probabilities of heroin use are logit linear in a random intercept u_i and some covariates and as well as the 'previous outcomes' $Y_{i,t-1}$:

$$\text{logit}[\Pr(Y_{it} = 1|Y_{i,t-1}, \boldsymbol{\beta})] = \eta_{it} = u_i + \beta_o + \beta_d d_{it} + \beta_t \ln t + \beta_{pv} Y_{i,t-1}.$$

TABLE 1. *Parameter estimates and s.e. (in italic) in models with and without ID (Int=Intercept, Prev=Previous, Pres=Present, Var=Variance).*

Model		Int	Dose	Time	Prev	Pres	Var
without ID	β	-0.643	-0.016	-0.421	1.430		1.837
		0.405	*0.006*	*0.072*	*0.140*		*0.414*
with ID	β	-0.640	-0.0164	-0.351	1.410		1.839
		0.386	*0.006*	*0.077*	*0.143*		*0.439*
	α	-6.460		0.695		2.039	
		0.986		*0.265*		*0.763*	

For the ID model, the conditional probabilities of drop-out are logit linear in some covariates including the 'present outcomes' $Y_{i,t}$ which signify ID:

$$\text{logit}[\Pr(R_{it} = 1|Y_{it}, \boldsymbol{\alpha})] = \zeta_{it} = \alpha_o + \alpha_t \ln t + \alpha_{ps} Y_{i,t}$$

for $t \leq n_i$ such that the present outcomes $Y_{i,t}$ are unobserved. When $t = n_i + 1$ and $n_i < 26$, the 'present outcome' Y_{i,n_i+1} is unobserved. Taking condition on the two possible values of Y_{i,n_i+1}, we have

$$\begin{aligned}&\text{logit}[\Pr(R_{i,n_i+1} = 1|Y_{i,n_i+1} = h, \boldsymbol{\alpha})] = \\ &\quad \zeta_{i,n_i+1,h} = \alpha_o + \alpha_t \ln t + \alpha_{ps} h, h = 0, 1.\end{aligned}$$

A vector of parameters for the whole model is $\boldsymbol{\theta} = (\boldsymbol{\beta'}, \boldsymbol{\alpha'})'$. Parameters of the model were estimated using Bayesian approach via Gibbs sampler and were implemented conveniently using WinBUGS package.

3 Results

We refer to Table 1 for a summary of our main results. Regarding the result of model with ID modeling, we found that reduced heroin use is significantly associated with increase in methadone dose and increase in duration of treatment. There is also a strong and positive association between the present and previous outcomes suggesting that some patients in treatment tend to use heroin continuously while others do not. Aparting from these covariates, the variance of the random intercepts is also significant showing the specificity of patients of the treatment. We even found that the positive and significant random intercepts help us to identify the heavy drug users of the programme. The significant parameters in the drop-out model, on the other hand, suggest that patients staying longer in the treatment are more likely to drop-out. Patients having take drugs are more likely to be absent for the coming urine test.

3.1 Discussion

Comparing the parameter estimates of models with and without informative drop-out modeling, they are more or less the same except that corresponding to time effect. The time effect (-0.351) in the drop-out modeling is decreased by 17% in magnitude or 7% in odds. This finding helps us to justify our concern on time effect. We are worrying the time effect of the treatment may be primarily due to the drop-out of heavy drug users which in turn leads to an impression that if patients stay longer in the treatment, they will reduce using drug. Now, we found that the time effect is indeed weaker after accounting for the drop-out process of the data which suggests some time effect may be because of the drop-out of heavy drug users.

References

Chan, J.S.K., Kuk, A.Y.C., Bell, J., and McGilchrist, C. (1998). The analysis of methadone clinic data using marginal and conditional logistic models with mixture or random effects. *The Australian & New Zealand Journal of Statistics*, **40**(1), 1–10.

Diggle P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics*, **43**, 49–93.

Little, R. J. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

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

Model Selection, Data Mining and Model Uncertainty

Chris Chatfield[1]

[1] Department of Mathematical Sciences, University of Bath, Bath, UK, BA2 7AY

Abstract: Different methods for selecting an appropriate model are briefly reviewed. Data mining, or data dredging, arises when large numbers of models are tried on the same data. The effects of this, such as model-selection bias, are still not widely understood and some remarks are made on model uncertainty.

Keywords: Akaike's information criterion; Bayesian information criterion; Data dredging; Principle of parsimony.

1 Introduction

Model-building is a complex procedure involving (i) the specification of an appropriate class of models; (ii) the selection of a 'best' model; (iii) model-fitting; and (iv) model-checking. In practice there are typically several cycles of fitting, because model-building is an *iterative, interactive* process as in the Box-Jenkins modelling of time series (Box et al, 1994). It may even be the case that more data are collected and new models entertained during the modelling cycle. Key considerations are the *objectives* of the study and the background knowledge - the *context*. In my work on *Pragmatic Statistical Inference*, I have explicitly included the context in the mathematical description of the problem (Chatfield, 2002). Another important decision is whether to use a black-box type of model or a structural model that accounts for specific physical features.
This paper concentrates on model selection and the related issues of data mining and model uncertainty.

2 Prelude

Consider the following 'typical' time-series problem. You have five years of monthly data and forecasts are required up to 12 months ahead. What would you do?
A standard approach would be to choose a family of possible models (e.g. the ARIMA class), look at the time plot and the sample autocorrelation function, try plausible models within the ARIMA family and choose a 'best'

model in some way. Then the analyst typically makes inferences and forecasts conditional on the selected model being 'true'. Statisticians all do this sort of thing but should they? The standard analysis ignores:

(1) The effects of Data Mining (DM). The model has been selected from the data.

(2) Model Uncertainty (MU). Uncertainty about the structure of the model is arguably the most important source of uncertainty, but also the least understood.

3 Model Selection

As well as finding out about any background knowledge, the analyst will typically begin by carrying out some form of *Initial Data Analysis* (e.g. Chatfield, 1995a, Chapter 6). Sometimes a model is selected subjectively using the results of this initial examination of the data, by using the analyst's experience to match an appropriate model to the observed characteristics. However, we concentrate here on the use of *model-selection criteria*. We cannot simply choose a model to give the best fit by minimizing the residual sum of squares, as this takes no account of the model complexity – the number of parameters fitted. There is an alternative fit statistic, called adjusted-R^2, which attempts to take account of the number of parameters, but more sophisticated model-selection statistics are generally preferred.
Akaike's Information Criterion (AIC) is the most commonly used and is given (approximately) by:

$$\text{AIC} \quad = \quad -2\ \ln(\text{max. likelihood}) + 2r,$$

where r denotes the number of independent parameters that are fitted for the model being assessed. Thus the AIC essentially chooses the model with the best fit, as measured by the likelihood function, subject to a penalty term, to prevent over-fitting, that increases with the number of parameters in the model. For an ARMA(p, q) model, note that $r = p + q + 1$ as the residual variance is included as a parameter. Ignoring arbitrary constants, the first (likelihood) term is usually approximated by $N \ln(S/N)$, where S denotes the residual sum of squares, and N is the number of observations. It turns out that the AIC is biased for small samples, and a bias-corrected version, denoted by AIC_C, is increasingly used. The latter is given (approximately) by replacing the quantity $2r$ in the ordinary AIC with the expression $2rN/(N - r - 1)$. The AIC_C is recommended, for example, by Brockwell and Davis (1991, Section 9.3) and Burnham and Anderson (2002).
An alternative, widely used, criterion is the **Bayesian Information Criterion** (BIC) that essentially replaces the term $2r$ in the AIC with the

expression $(r + r \ln N)$. This penalizes the addition of extra parameters more severely than the AIC.
Several other possible criteria have also been proposed - see Burnham and Anderson (2002) for a general review. Note that all the criteria may not have a unique minimum and depend on assuming that the data are (approximately) normally distributed. Although dimensionless, the arithmetic value of a statistic like the AIC is hard to interpret unless it is first subtracted from the value for the 'best' model. A rule of thumb is that models having an AIC within 2 or 3 of the minimum value are 'good', while models with differences greater than about 8 should be discarded. Following the results in Faraway and Chatfield (1998), I generally prefer to use the AIC_C or BIC. Computer packages routinely produce numerical values for several such criteria so that the analyst can pick the one he or she likes best. The guiding principle throughout is to apply the Principle of Parsimony, which says '*Adopt the simplest acceptable model*'.
An alternative approach to model selection relies on carrying out a series of *hypothesis tests*. Econometricians tend to favour this approach and may test null hypotheses for such attributes as normality, constant variance, unit roots and non-linearity. However, little will be said here about this approach, because the author, like most statisticians, prefers to rely on the subjective interpretation of diagnostic tools (such as the correlogram for time-series data) allied to the model-selection criteria given above. My reasons for this preference are as follows:

1. A model-selection criterion gives a numerical-valued ranking of all models, so that the analyst can see if there is a clear winner or, alternatively, if there are several close competing models.

2. It is difficult to use hypothesis tests to compare non-nested models.

3. A hypothesis test requires the specification of an appropriate null hypothesis, and effectively assumes the existence of a true model that is contained in the set of candidate models.

Of course, there is a real danger that the analyst will try many different models, pick the one that appears to fit best according to one of these criteria, but then make predictions as if certain that the best-fit model is the true model. Further remarks on this problem are made in Section 4.

4 Data Mining

What exactly is **Data Mining** (DM)? Is it 'good' or 'bad'? Is it 'getting as much as you can out of a set of data', or 'squeezing your data dry and perhaps finding spurious relationships'? Typically analysts try several models, pick the 'best' one, and then behave as if this was the only model fitted. This is dangerous. For example, in multiple regression, the analyst

may try p explanatory variables and select $q < p$ 'significant' variables. If we then take $(q + 1)$ as the model D.F., this disregards the fact that q itself has been estimated from the data. As a result, the variance of out-of-sample predictions will be under-estimated. There are similar problems in hypothesis-testing when a hypothesis is generated and tested on the same data. Problems are likely to be most serious for observational data (e.g. time series) and least serious for designed experiments with prior hypotheses (e.g. clinical trials).

The term DM has been around for many years and, in particular, is used in the econometrics literature to mean data-dependent specification searches. An alternative description is **Data-Dredging**, and this may be preferable because the term DM is also used by computer scientists to denote the very different activity involved in extracting previously unknown and potentially useful information from databases that may be large, noisy and have missing data. This form of DM is sometimes called **Knowledge Discovery in Databases** or KDD. Techniques used here include various classification tools, neural nets, genetic algorithms, and clustering methods – see for example Hand (1998) and Hastie et al (2001). Computing hype makes grandiose claims, which are not always borne out by results. There are substantial differences from DM as used in statistical model-selection. For computer science DM, the data have often been collected electronically for some other purpose and are often non-numerical. Datasets are very, very large (meaning millions+). It is no longer possible to 'look' at all the data, and any analysis has to be automated. It is more usual to 'apply an algorithm' rather than 'fit a model'. Small differences may be statistically significant, but are they of practical importance? Is a significance test valid anyway? Local patterns are of interest, but how do we distinguish those that arise by chance? (e.g. Fraud detection looks for unusual patterns.) Clearly, there is much of interest here for statisticians.

Returning to Statistical DM, or Data-Dredging, we note that it is often seen as a rather suspect activity. However, this should really depend on how it is applied. A possible way to describe a 'good' version of DM is: Trying many different models on a set of data as a way of *generating* hypotheses. However, a 'bad' version of DM is to try many different models on a set of data and then behave as if the 'best' model is true and the only one that has been fitted.

Surprisingly little seems to be known about the effects of data-dependent model specification (Chatfield, 1995b). Data dredging yields models which overfit the data and may be poor at out-of-sample predictions. Likewise tests on hypotheses generated by the data are likely to be 'significant' when the same data are used for the test.

5 Model Uncertainty

When building a model, it is easy to forget that there is probably no such thing as a 'true' model (except in simulation exercises), and that all models are *tentative* and *approximate*. Yet statistical theory typically assumes we know the true model.

Nowadays computers let us look quickly at tens, or even hundreds of models. Thus, we effectively admit MU by searching for the 'best' model but then ignore MU when making inferences. If we formulate and fit model to the same data (the usual situation!), it is easy to show that standard theory does *not* apply. There can be large model-selection biases, especially in time-series modelling, and in multiple regression (Chatfield, 1995b). Statisticians are much more familiar with two other sources of uncertainty, namely parameter uncertainty, assuming the model is known, and unexplained random variation. In my experience, uncertainty arising from estimating parameters is usually much less important than uncertainty about the model structure, but gets far more attention. Structural uncertainty can arise, not only through incorrect model selection, but also because the underlying model is changing through time, or because there is no 'true' model anyway. One immediate consequence of model uncertainty is that out-of-sample forecasts generally have poorer accuracy than expected from within-sample fit.

Some alternatives to trying to pick a 'best' model are to combine several plausible models, to fit different models to different parts of data or to use a model whose parameters are allowed to adapt through time. Further details on the effects of model uncertainty, and ways of dealing with it, are given by Chatfield (1995b; 1996; 2001, Chapter 8).

6 Postscript

My most recent large-scale modelling exercise (Zidek et al, 2003), on exposure to air particles (PM_{10}), has been very different to my previous experience in time-series analysis. Much of the model was pre-specified from environmental considerations and the major effort was in collecting, and handling fairly large datsets (for example one year's hourly observations gives 8760 observations). From bitter experience, here are some tips. (1) There *will* be data peculiarities; (2) Check the first and last lines of the data carefully; (3) Check there are the right number of lines of data; (4) Check the effects of changing to summertime - if one day has 25 observations, this can ruin the analysis! Dealing with practical modelling problems like this is at least as important as understanding relevant theory.

References

Box, G.E.P., Jenkins, G. M., and Reinsel, G.C. (1994). *Time Series Analysis, Forecasting and Control*, 3rd edn. Englewood Cliffs, NJ: Prentice-Hall.

Brockwell, P.J. and Davis, R.A. (1991). *Time Series: Theory and Methods*, 2nd edn. New York: Springer-Verlag.

Burnham, K.P. and Anderson, D.R. (2002). *Model Selection and Multi-Model Inference*, 2nd edn. New York: Springer-Verlag.

Chatfield, C. (1995a). *Problem-Solving: A Statistician's Guide*, 2nd edn. London: Chapman & Hall.

Chatfield, C. (1995b). Model uncertainty, data mining and statistical inference (with discussion). *Journal of the Royal Statistical Society, Series A*, **158**, 419-466.

Chatfield, C. (1996). Model uncertainty and forecast accuracy. *Journal of Forecasting*, **15**, 495–508.

Chatfield, C. (2001). *Time-Series Forecasting*. Boca Raton: Chapman & Hall/CRC Press.

Chatfield, C. (2002). Confessions of a pragmatic statistician. *The Statistician*, **51**, 1–20.

Faraway, J. and Chatfield, C. (1998). Time series forecasting with neural networks: A comparative study using the airline data. *Applied Statistics*, **47**, 231–250.

Hand, D.J. (1998). Data Mining: Statistics and more? *American Statistician*, **52**, 112–8.

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

Zidek, J.V., Meloche, J., Shaddick, G.S., Chatfield, C., and White, R. (2003). A computational model for estimating personal exposure to air pollutants with application to London's PM_{10} in 1997. Technical Report no. 2003-3, Statistical and Applied Mathematical Sciences Institute, RTP, North Carolina.

Frequentist Model Averaging and Model Selection

Gerda Claeskens[1] and Nils Lid Hjort[2]

[1] Department of Statistics, Texas A&M University, 447 Blocker Building, College Station, TX-77843, USA. Email: Gerda@stat.tamu.edu.
[2] Department of Mathematics, University of Oslo, P.O. Box 1053 Blindern, N-0316 Oslo, Norway

Abstract: Frequentist model averaging estimators are studied using large-sample likelihood methodology. Limiting distributions of post-model selection estimators and of estimators averaged across models are obtained. A study of limiting risk of the estimators is performed. An unbiased estimator of the limiting risk leads to a focussed information criterion, the FIC, for model selection.

Keywords: Model averaging estimators; Model information criteria; Variable selection.

1 Motivation

The traditional use of model selection methods in practice is to proceed as if the final selected model had been chosen a priori, without acknowledging the additional uncertainty introduced by the model selection step. As a consequence, coverage probabilities of confidence intervals conditional on the selected model will often be much smaller than the nominal coverage probabilities and variability might be underreported.
We discuss the topic of post-model selection and of frequentist model averaging estimators and their asymptotic distributions. Existing results are extended in several directions. We obtain risk properties of estimators-post-model selection as well as of estimators averaged across models, and take modelling bias explicitly into account. Our methodology is applicable to any model selection mechanism and to general modelling settings, which include regression models and generalized linear models.
The model averaging work naturally leads to a new class of model selection criteria which put special focus on the parameter singled out for inference. This leads to a focussed or concentrated information criterion.
For more information and details we refer to Hjort and Claeskens (2003) and Claeskens and Hjort (2003).

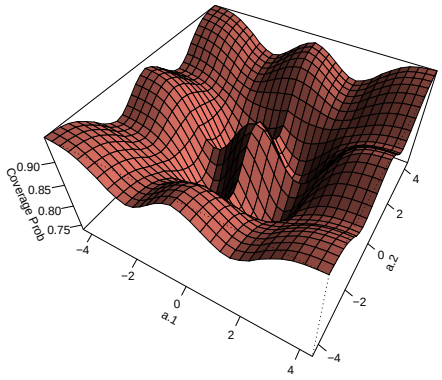


FIGURE 1. *Coverage probability when ignoring a model selection step in AIC model choice between four models, as a function of location in a local alternative model space. The nominal coverage probability is 95%, while the minimum obtained coverage value equals 0.748.*

2 Post-model Selection Estimators

The need to study the asymptotic behaviour of estimators after model selection is clearly illustrated when considering coverage probabilities. Figure 1 shows the coverage deficiency when ignoring a model selection step in further inference for Akaike's information criterion AIC choosing between four models: model (1) is the minimal model, models (2) and (3) each add one extra variable to the minimal model and model (4) adds both of these variables. The intended nominal coverage probability is 95%.
The key to derive properties of post-model selection estimators is that these estimators can be considered as special cases of estimators averaged across models. Consider a collection of possible models. Using the data at hand, a model selector will pick a single one of these models, for example that model with best predictive qualities or smallest distance to the in some sense true model. We can write the post-model selection estimator as a weighted average of the estimator of μ in the different models under consideration, where the random weight function is one if this model is the one selected by our data-dependent selection criterion, and is zero otherwise. Particularities of the specific mechanism, such as for example AIC or BIC, are contained in this indicator function. Post-model selection estimators are a substantial motivation to study frequentist model averaging estimators in general.

3 Frequentist Model Averaging

When interest is in finding a best parameter estimate, and not as much in model selection, an interesting area is frequentist model averaging. A general class of model averaging estimators weights estimators in different models by, typically, data dependent weights, summing to one. Note that the class of model average estimators includes the post-model selection estimators as a special case. A general model averaging estimator, however, bypasses the model selection step.

From a frequentist perspective, not many results are known about model averaging, however, see Hjort and Claeskens (2003). This is in contrast to the Bayesian literature, where many methods have been discussed, partly focussing on algorithmic matters. In Bayesian model averaging, priors are put on all of the models under consideration. Here we take a purely non-Bayesian approach to the problem.

As a main result we obtain the limiting distribution of the model averaging estimators, which is a suitable convex mixture of Gaussian random variables, of which the mixing coefficients are largely determined by the choice of weights. The asymptotic distribution of the estimators is obtained in the local misspecification framework where the true parameter is at a distance $1/\sqrt{n}$ from the parameter used in our models. This encompasses for each subset model the possibility that this particular submodel is correct and the others incorrect. We obtain expressions for the limiting mean squared error of the model average estimators in this setting. The limiting distribution of model averaging estimators also leads us to expressions for the real coverage probability of confidence intervals when the model uncertainty is ignored, this to learn about how much can be lost in the traditional approach. We used this methodology to obtain Figure 1 presented above.

Since the correct limiting distribution is non-normal, we need formulae for constructing asymptotically correct confidence intervals. These formulae can be derived from the limiting distribution, see Hjort and Claeskens (2003) for details.

4 Parameter Adaptive Model Selection

Asymptotic expressions for risk functions of model averaging and post-model selection estimators are a function of the estimated parameter. This poses the question of whether model selection criteria can and should be made parameter adaptive. While model selection criteria such as AIC and BIC, along with several variations, aim at selecting a single model with good overall properties, none of these methods is concerned with the actual use of the selected model, which varies with application and context. A model which gives good precision for one estimand might be worse when used for another estimand. This indicates that improvement might be possible if focus is restricted to the selection of a model for a specific parameter.

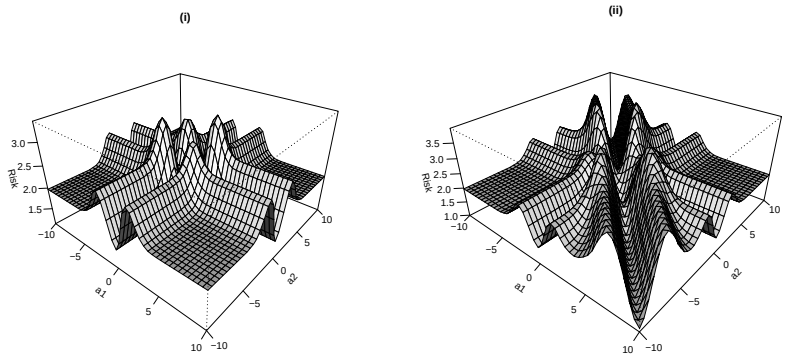


FIGURE 2. *Asymptotic risk of the model selection criteria (i) AIC and (ii) FIC, the focussed information criterion, for model selection amongst four models. Risk is depicted as a function of location in the local alternative model space.*

We construct an unbiased estimator of the asymptotic risk expression which leads to a new type of model selection criterion which focusses on the parameter of interest. The framework is that of large-sample likelihood inference. We investigate properties of this new criterion and discuss some connections to the AIC. Figure 2 shows the different behaviour in risk for (i) AIC model selection amongst 4 models (see earlier) and (ii) selection using the focussed information criterion FIC.

5 Applications

A first application is averaging over logistic regression models. The data contain information on factors that might influence the birth weight of babies. The classical AIC and BIC criteria do not select the same model for these data. Interest is in estimating the probability of low birth weight for babies of the average 'white' and 'black' mothers, and for the ratio of these two probabilities. The focussed information criterion is applied to select a model for each of these parameters of interest and suggests different variables to be included for the different parameters. Note that both AIC and BIC provide only one best model, ignoring the focus parameter. Further we apply several model averaging strategies to estimate these parameters along with standard deviations and confidence bounds.
A second application consists of averaging over covariance structure models. We use data from the Adelskalenderen of speedskating, which lists the best speedskaters ever, as ranked by their personal best times over the four distances 500, 1500, 5000 and 10000 m. The correlation structure of the 4-vector of times is important when relating performances on the different

distances. We illustrate how models for the covariance structure are averaged to form estimators of quantities of interest. Also here, our theory provides methodology to obtain standard errors and confidence intervals.

References

Claeskens, G. and Hjort, N.L. (2003). The focussed information criterion. *Journal of the American Statistical Association*. (in press).

Hjort, N.L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*. (in press).

A Version of the EM Algorithm for Proportional Hazards Model with Random Effects

José Cortiñas Abrahantes[1] and Tomasz Burzykowski[1]

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

Abstract: Proportional hazards models with random effects (frailties) have been the focus of the research on methods of analysis of multivariate failure-time data. Several estimating methods have been proposed to tackle the problem of the estimation of the parameters (Xue and Brookmeyer, 1996; Vaida and Xu, 2000; Ripatti et al, 2002). An alternative fitting approach based on a modificated EM algorithm, in which the Laplace approximation is used at the E-step is presented in this paper. The performance of the method is assessed based on a simulation study.

Keywords: Frailty model; Laplace approximation; Multivariate failure-time data.

1 Introduction

Proportional hazards models with random effects (frailties) acting multiplicatively on the baseline hazard have been the focus of the research on methods of analysis of multivariate failure-time data for a long time. Initially, the research concentrated on univariate (shared) frailty models. The models have some limitations. Therefore, multivariate frailty models have started to attract some attention. Several estimating methods have been proposed to tackle the problem of the estimation of the parameters (Xue and Brookmeyer, 1996; Vaida and Xu, 2000; Ripatti et al, 2002). In this paper an alternative fitting approach is considered. It is based on a modification EM algorithm, in which the Laplace approximation is used at the E-step. We consider clustered failure-time data with N clusters. The failure-time variable corresponding to subject j ($i = 1, \ldots, n_i$) from cluster i ($i = 1, \ldots, N$) is denoted by T_{ij}. It is assumed that observations of T_{ij} can be right-censored. Thus, for subject j in cluster i, we observe $Y_{ij} = \min(C_{ij}, T_{ij})$, where C_{ij} is a random censoring time independent of T_{ij}. Additionally, a censoring indicator δ_{ij} is observed, with δ_{ij} equal to 1 if $Y_{ij} = T_{ij}$, and 0 if $Y_{ij} = C_{ij}$. The following mixed-effects proportional hazards model for T_{ij} is considered:

$$\lambda(t_{ij}|\beta_i, b_i) = \lambda_0(t_{ij}) \exp(x_{ij}^T \beta_i + z_{ij}^T b_i),$$

where $\lambda_0(t)$ is the baseline hazard function, β_i is a vector of cluster-specific fixed-effects corresponding to a vector of covariates x_{ij}, and b_i is a vector of random effects associated with a vector of covariates z_{ij}. The b_i are assumed randomly distributed with mean 0 and variance-covariance matrix $D = D(\theta)$. The density function of b_i is denoted by $f(t)$. Its specification is not necessary at this point; later on we will assume that it corresponds to a mean-zero multivariate normal distribution.
The estimation of the parameters β_i and θ from the observed data on T_{ij} is our main interest. Assuming the conditional independence of the observations given b_i, one might write the (conditional) log-likelihood for the observed data in the ith cluster:

$$l_i(\beta_i, \lambda_0|b_i) = \sum_{j=1}^{n_i} \{\delta_{ij}[\log \lambda_0(t_{ij}) + x_{ij}^T\beta_i + z_{ij}^T b_i] - \Lambda_0(t_{ij}) \exp(x_{ij}^T\beta_i + z_{ij}^T b_i)\}.$$

The (marginal) likelihood of the observed data for all clusters can then be expressed as follows:

$$L(\beta_i, \theta, \lambda_0) = \prod_{i=1}^{N} \int L_i^\star(\beta_i, \theta, \lambda_0, b_i) db_i, \tag{1}$$

where

$$L_i^\star(\beta_i, \theta, \lambda_0, b_i) = \prod_{i=1}^{n_i} e^{l_i(\beta_i, \lambda_0|b_i)} f(b_i, D(\theta)). \tag{2}$$

Note that (2) can be treated as the likelihood of the "augmented" data for cluster i, treating b_i as additional observations. Consequently,

$$L^\star(\beta, \theta, \lambda_0, b) = \prod_{i=1}^{N} L_i^\star(\beta_i, \theta, \lambda_0, b_i), \tag{3}$$

is the likelihood of the "augmented" data for all clusters, with β and b denoting vectors resulting from "stacking" vectors β_i and b_i, respectively, for all clusters.

2 The EM Algorithm and Issues in the Implementation

The EM algorithm consists of two steps: the E-step and the M-step. Starting from initial values of parameters β_i, θ and λ_0, the algorithm iterates between the E-step, and the M-step. The algorithm is iterated until convergence is reached.
At the E-step the expectation of the logarithm of the likelihood (3), conditional on the observed data and the current values $\widetilde{\beta}_i$, $\widetilde{\theta}$ and $\widetilde{\lambda}_0$ of parameters β_i, θ and λ_0, is required. The expectation will be denoted by

$Q(\beta_i, \theta, \lambda_0)$. It appears that it can be expressed as:

$$Q(\beta_i, \theta, \lambda_0) = Q_1(\beta_i, \lambda_0) + Q_2(\theta), \tag{4}$$

where $Q_1(\beta_i, \lambda_0)$ is equal to:

$$\sum_{i=1}^{N}\sum_{j=1}^{n_i}\{\delta_{ij}[\log\lambda_0(t_{ij}) + x_{ij}^T\beta_i + z_{ij}^T\mathrm{E}(b_i)] - \Lambda_0(t_{ij})\exp[x_{ij}^T\beta_i + \log\mathrm{E}(e^{z_{ij}^T b_i})]\}$$

and

$$Q_2(\theta) = \sum_{i=1}^{N}\mathrm{E}[\log f(b_i)],$$

with E(.) denoting expected values (conditional on the observed data and the current values of the parameters).
In the M-step, new estimates $\widetilde{\beta}_i$ and $\widetilde{\theta}$ are found by maximizing the functions Q_1 and Q_2, respectively. The use of the EM algorithm, as described above, is complicated by the need to compute the conditional expected values at the E-step. The expectations involve integrals of the form

$$\mathrm{E}\{g(b_i)\} = \frac{\int g(b_i)e^{l_i(\tilde{\beta}_i,\tilde{\lambda}_0|b_i)+\log f(b_i;D(\tilde{\theta}))}db_i}{\int e^{l_i(\tilde{\beta}_i,\tilde{\lambda}_0|b_i)+\log f(b_i;D(\tilde{\theta}))}db_i}\ . \tag{5}$$

Usually, they will not be available in a closed-form. To compute them, Xue and Brookmeyer (1996) proposed to use numerical integration. Vaida and Xu (2000) and Ripatti et al (2002) proposed to use MCMC methods. An alternative solution, not yet considered in the literature, is to use the Laplace approximation (Evans and Swartz 2000). Using the Laplace approximation, it can be shown that the integrals (5) can be approximated by

$$\mathrm{E}\{g(b_i)\} \approx g(\hat{b}_i)\ , \tag{6}$$

where $\hat{b}_i$ is an isolated global maximum of

$$K(b_i) = -\frac{1}{n_i}[l_i(\tilde{\beta}_i, \tilde{\lambda}_0|b_i) + \log f\{b_i; D(\tilde{\theta})\}]\ . \tag{7}$$

Though computing the approximation requires finding the maximum of the function $K()$, it is numerically less demanding than, e.g., numerical integration or MCMC sampling.
The variance covariance matrix of the solution $(\hat{\beta}_i, \hat{\lambda}_0, \hat{\theta})$ obtained from the EM algorithm, can be estimated using the inverse of the observed Fisher information matrix. The latter matrix can be computed by the formula developed by Louis (1982).

3 Application and Simulation Study

A practical application is considered in the context of the validation of surrogate endpoints. Two datasets, coming from multiple randomized cancer clinical trials, are analyzed. The first one includes data from two multicenter trials in advanced colorectal cancer, aiming at the evaluation of the benefits of experimental fluoropyrimidine treatments vs. the use of 5-fluorouracil (5FU). The second data set contains four randomized multicenter trials in advanced ovarian cancer. The trials compared the treatment with cyclosphosphamide plus cisplatin vs. the treatment with cyclosphosphamide plus adriamycin plus cisplatin. To assess the validity of disease-free survival time (S_{ij}) as a surrogate for overall survival time (T_{ij}), the following proportional hazards model, with center-specific, random treatment effects b_{Si} and b_{Ti} and unspecified baseline hazards λ_{Si} and λ_{Ti} (stratified by center), is fitted to each of the data sets:

$$\lambda_{Sij}(s_{ij}|\beta_{Si}, b_{Si}) = \lambda_{Si}(s_{ij})e^{x_{ij}^T\beta_{Si}+Z_{ij}^T b_{Si}}\,, \tag{8}$$

$$\lambda_{Tij}(t_{ij}|\beta_{Ti}, b_{Ti}) = \lambda_{Ti}(t_{ij})e^{x_{ij}^T\beta_{Ti}+Z_{ij}^T b_{Ti}}\,. \tag{9}$$

Note that in (8)–(9) i and j denote the center and the patient, respectively, and Z_{ij} is a binary covariate indicating the treatment group. The random treatment effects b_{Si} and b_{Ti} are assumed to follow a mean-zero bivariate normal distribution. The parameter of interest is the square of the correlation between the random treatment effects, as it is related to the precision of the prediction of the treatment effect on the true endpoint T from the effect on the surrogate S. The results obtained from model (8)–(9) are compared to those obtained by Burzykowski et al (2001) using copula models. In general, the point estimates of the parameter of interest are comparable, while the estimates of the standard error are smaller for model (8)–(9).
A simulation study, in which clustered bivariate failure-time data are generated under a model similar to (8)–(9), but with non-stratified baseline hazards and correlated random intercepts instead of the random covariate effects. Several configurations of the parameters of the simulation model are considered, allowing for the investigation of the performance of the proposed estimation method in function of the number of clusters, the number of subjects per cluster, the percentage of censored observations, the variances and the correlation associated with the random effects. The following results of the simulations, pertaining to the estimation of the variance-covariance structure of the random effects, are observed:

- the correlation between the random intercepts is underestimated, while their variances are overestimated;
- the absolute relative bias in the estimates of the variances and the correlation depends on the cluster size n_i: it is around $20-30\%$ when $n_i = 20$, and drops below 10% when $n_i \geq 50$;

- there is almost no effect of the number of clusters N on the absolute relative bias of the estimates of the correlation;
- the absolute relative bias of the estimates for the variances decreases with N; for $N = 10, 20$ underestimation is observed, while for $N = 50, 100$ an overestimation appears.
- the absolute relative bias for the correlation increases with the magnitude of the correlation;
- 20% of censoring slightly increases the bias;
- in general, the use of the observed Fisher information matrix, computed by the method of Louis (1982), leads to only a slight underestimation of the standard errors of the estimated parameters.

4 Concluding Remarks

In summary, one may conclude that the proposed method of the estimation of proportional hazards models with random effects does offer an advantage in terms of the numerical complexity, as compared to the other proposals based on the EM algorithm (Xue and Brookmeyer, 1996; Vaida and Xu, 2000; Ripatti et al, 2002). Due to the asymptotic nature of the Laplace approximation, however, the proposed method does require sufficient amount of data per cluster to provide reasonable results.

References

Burzykowski, T., Molenberghs, G., Buyse, M., Geys, H., and Renard, D. (2001). Validation of surrogate endpoints in multiple randomized clinical trials with failure-time endpoints.*Journal of the Royal Statistical Society C (Applied Statistics)*, **50**, 405–422.

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

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **44**(2), 190–200.

Ripatti, S., Larsen, K., and Palmgren, J. (2002). Maximum likelihood inference for multivariate frailty models using an automated Monte Carlo EM algorithm. *Lifetime Data Analysis*, **8**, 349–360.

Vaida, F. and Xu, R. (2000). Proportional hazards model with random effects. *Statistics in Medicine*, **19**, 3309–3324.

Xue, X. and Brookmeyer, R. (1996). Bivariate frailty model for the analysis of multivariate survival time. *Lifetime Data Analysis*, **2**, 277–289.

Using P-splines to Extrapolate Two-dimensional Poisson Data

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: Eilers & Marx (1996) used P-splines to smooth one-dimensional count data with Poisson errors. In this paper we consider the extrapolation problem and show that P-splines are well suited to extrapolating in both one and two dimensions. The role of the order of the penalty is highlighted. We illustrate our remarks with the analysis of a large set of mortality data indexed by age of death and year of death.

Keywords: Mortality; P-splines; Extrapolation; Smoothing; Two-dimensions.

1 Introduction

The method of P-splines (Eilers and Marx, 1996) is now well established as a method of smoothing in generalized linear models (GLMs). A succinct summary of the method is: (a) use B-splines as the basis for the regression, and (b) modify the log-likelihood by a difference penalty on the regression coefficients. Wand (2003) gives a most useful overview which highlights the wide class of models that can be fitted with the P-spline approach.
Durban, et al (2002) introduced a two-dimensional P-spline model for Poisson data in which the regression matrix was defined in terms of the Kronecker product of the regression matrices of two one-dimensional P-spline models. The present paper shows that P-splines provide a natural method of extrapolating the fitted mortality rates forward in time. The role of the order of the penalty is shown to be of particular importance. We illustrate our remarks with the analysis of the same set of mortality data as our 2002 paper.

2 Description of the Data

The failure to predict accurately the fall in UK mortality rates from the 1970s to date has had far-reaching consequences for the pensions and annuity business of the UK insurance industry. The Continuous Mortality Investigation Bureau (CMIB) has responsibility for monitoring and predicting mortality rates. In this paper we consider one of the CMIB data sets, namely that for male assured lives. For each calendar year (1947 to 1999) and each age (11 to 100) we have the number of years lived (the exposure) and the number of policy claims (deaths). We use a Kronecker product P-spline model (Durban, et al., 2002) and a system of prior weights to predict mortality rates for 1975-1999 using the data from 1947-1974. The comparison between the observed rates for 1975-1999 and our predicted rates provides a good test of our method.

3 Extrapolating Mortality Tables

Our data consists of two matrices, $\boldsymbol{Y}$ and $\boldsymbol{E}$, whose rows are indexed by age (here 11 to 100) and whose columns are indexed by year (here 1947 to 1999). The matrix $\boldsymbol{Y}$ contains the number of claims (deaths) and the matrix $\boldsymbol{E}$ contains the exposures. Thus $\boldsymbol{R} = \log(\boldsymbol{Y}/\boldsymbol{E})$ is the matrix of raw log hazards. Durban, et al (2002) showed how to smooth $\boldsymbol{R}$ by using a 2-dimensional extension of the P-spline model of Eilers and Marx (1996). The smoothing is achieved by using a penalized generalized linear model (PGLM) for $\boldsymbol{Y}$ with Poisson errors and appropriately defined regression and penalty matrices.

We define the regression matrix in terms of the Kronecker product of two 1-dimensional regression matrices. Let $\boldsymbol{B}_a = \boldsymbol{B}(\boldsymbol{x}_a)$, $n_a \times c_a$, be a regression matrix of B-splines based on the explanatory variable $\boldsymbol{x}_a$; in our example, $\boldsymbol{x}_a' = (11, \ldots, 100)$ so $n_a = 90$ and c_a is typically about 20. Similarly, let $\boldsymbol{B}_y = \boldsymbol{B}(\boldsymbol{x}_y)$, $n_y \times c_y$, be a regression matrix of B-splines based on the explanatory variable $\boldsymbol{x}_y$; in our example, $\boldsymbol{x}_y' = (1947, \ldots, 1999)$ so $n_y = 53$ and c_y is typically about 10. The regression matrix for our 2-dimensional model is the Kronecker product

$$\boldsymbol{B} = \boldsymbol{B}_y \otimes \boldsymbol{B}_a. \tag{1}$$

This formulation assumes that the vector of observed claim numbers $\boldsymbol{y} = \mathrm{vec}(\boldsymbol{Y})$, (this corresponds to how Splus stores a matrix). Note that $\boldsymbol{B}$ has $n_a n_y$ rows and $c_a c_y$ columns, so is typically 4770 by 200. The model is, at present, a standard GLM: $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$ where $\log \boldsymbol{\mu} = \log \boldsymbol{e} + \boldsymbol{B}\boldsymbol{a}$ and $\log \boldsymbol{e}$, $\boldsymbol{e} = \mathrm{vec}(\boldsymbol{E})$, is the usual offset in a log linear model for mortality data.

This regression model will usually be over-parameterized ($\mathrm{len}\,(\boldsymbol{a}) \approx 200$) so we introduce a penalty on $\boldsymbol{a}$. (Durban, et al, 2002) show that an appropriate

penalty matrix is

$$\boldsymbol{P} = \lambda_a \boldsymbol{I}_{c_y} \otimes \boldsymbol{D}_a' \boldsymbol{D}_a + \lambda_y \boldsymbol{D}_y' \boldsymbol{D}_y \otimes \boldsymbol{I}_{c_a} \tag{2}$$

where $\boldsymbol{I}_{c_a}$ is an identity matrix of size c_a and $\boldsymbol{D}_a$ is a difference matrix with dimension $(c_a - p_a) \times c_a$ where p_a is the order of the penalty on age; similar definitions apply for $\boldsymbol{I}_{c_y}$ and $\boldsymbol{D}_y$. For given values of the smoothing parameters λ_a and λ_y the model is fitted by penalized likelihood and the penalized version of the scoring algorithm

$$(\boldsymbol{B}'\tilde{\boldsymbol{W}}\boldsymbol{B} + \boldsymbol{P})\hat{\boldsymbol{a}} = \boldsymbol{B}'\tilde{\boldsymbol{W}}\boldsymbol{B}\tilde{\boldsymbol{a}} + \boldsymbol{B}'(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}). \tag{3}$$

Here, $\tilde{\boldsymbol{a}}$, $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{W}} = \text{diag}(\tilde{\boldsymbol{\mu}})$, the diagonal matrix of weights, denote current estimates, and $\hat{\boldsymbol{a}}$ denotes the updated estimate of $\boldsymbol{a}$; additionally, $\log \boldsymbol{\mu} = \log \boldsymbol{e} + \boldsymbol{B}\boldsymbol{a}$, the canonical link. Finally, the smoothing parameters can be selected by optimising with respect to the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), for example. We perform extrapolation with the following simple device: we define a weight matrix $\boldsymbol{V} = \text{blockdiag}\{\boldsymbol{I}, \boldsymbol{0}\}$ where $\boldsymbol{I}$ is an identity matrix of size $n_a n_{y_1}$ and $\boldsymbol{0}$ is a square matrix of 0's of size $n_a(n_y - n_{y_1})$. We have in mind using n_{y_1} years of data as a training set and extrapolating the remaining $n_y - n_{y_1}$ years. Alternatively, we can take $\boldsymbol{I}$ to have size $n_a n_y$ and extrapolate into the future. To accommodate the weight matrix $\boldsymbol{V}$ we modify the scoring algorithm (3) as follows:

$$(\boldsymbol{B}'\boldsymbol{V}\tilde{\boldsymbol{W}}\boldsymbol{B} + \boldsymbol{P})\hat{\boldsymbol{a}} = \boldsymbol{B}'\boldsymbol{V}\tilde{\boldsymbol{W}}\boldsymbol{B}\tilde{\boldsymbol{a}} + \boldsymbol{B}'\boldsymbol{V}(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}) \tag{4}$$

where any unknown values in $\boldsymbol{y}$ and $\boldsymbol{e}$ can be given arbitrary values.
Example: We illustrate our methodology by using the 1947-1974 data to predict the 1975-1999 rates. Figure 1 shows the fitted and extrapolated log(mortality) values for ages 35 and 60. The fit used cubic B-splines and second order difference penalties; the smoothing parameters were chosen using BIC. Confidence intervals are also included and we note that the observed rates for 1975-1999 for both ages are comfortably within their respective 95% confidence funnels.

4 The Role of the Order of the Penalty

In the previous section we used a quadratic penalty, $p_a = p_y = 2$. In this section we examine the conventional wisdom that the order of the penalty has only a small effect on any smoothed values. Figure 2 shows the results of fitting and extrapolating using first order ($p_a = p_y = 1$), second order ($p_a = p_y = 2$) and third order penalties ($p_a = p_y = 3$). We make two comments: first, the order of the penalty has no discernible effect on the smooth of the training data; second, the order of the penalty has a dramatic effect on the extrapolated values. In this paper we have concentrated on

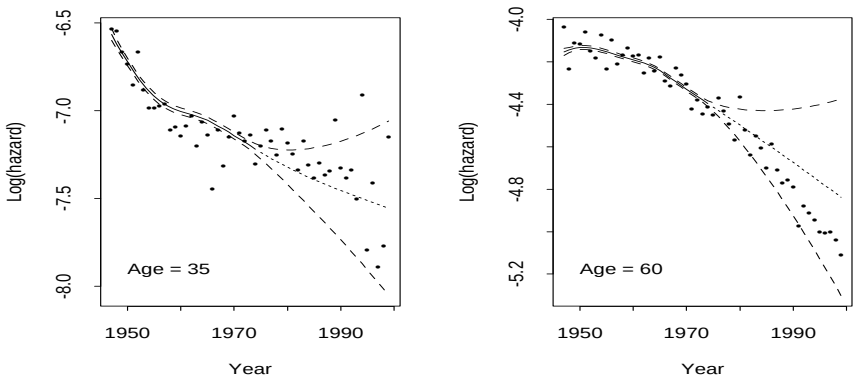


FIGURE 1. *Observed, fitted and extrapolated log(hazard) with 95% confidence intervals for* $p_a = p_y = 2$. *Left panel: age 35, right panel: age 60.*

the 2-dimensional problem but it is clear from (4) that the method can be applied in 1-dimension. In this case it can be shown that the extrapolation works by extrapolating the regression coefficients and these extrapolations are constant, linear or quadratic depending on the order of the penalty. This result is approximately true in 2-dimensions, as is evident from Figure 2. We make some further comments on this property in our concluding remarks.

5 Conclusions

The failure to predict accurately the fall in mortality rates has had far-reaching consequences for the UK pensions and annuity business. What comfort can be drawn from the results presented in this paper? We compare the predicted mortality rates from 1975-1999 with the observed rates over the same period and draw two main conclusions.

First, the predicted rates are higher than the observed rates for nearly all ages. Visual inspection of the observed rates suggests that it is unlikely that the sharp fall in mortality that occurred from the 1970's to the present could have been predicted back in the 70's.

Second, from 1975 to date, the observed rates lie at about one standard error below the predicted rates and are comfortably within the confidence funnel of the predicted rates. In view of the variation in the mortality rates observed before 1975 this suggests that a prudent course is to allow for this variation by discounting the predicted rates by a certain amount. If this discount had been set at one standard error then the resulting 'prudent'

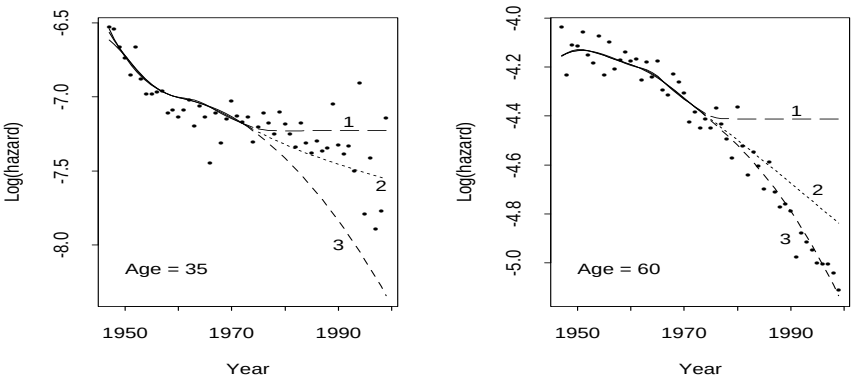


FIGURE 2. *Observed, fitted and extrapolated log(hazard) for* $p_a = p_y = 1$*, 2 and 3 in turn. Left panel: age 35, right panel: age 60.*

predictions would have been very close to what actually happened. Our view is that some such discounting procedure is the only reasonable way of allowing for the uncertainty in these, or indeed any, predictions.

We also make two general remarks on our method. First, we emphasise the critical role of the order of the penalty, *pord*. The choice of the order of the penalty corresponds to a view of the future pattern of mortality: *pord* = 1, 2 or 3 corresponds respectively to future mortality continuing at a constant level, improving at a constant rate or improving at an accelerating (quadratic) rate. We not only used BIC to choose the values of the smoothing parameters for given value of *pord* we also used BIC to choose the value of *pord*; the preferred value of *pord* was 2 and this was used to produce Figure 1.

Second, in this paper we have been concerned with extrapolation forward in time. However, the method is quite general. In one dimension we can extrapolate both forward and backward while in two dimensions we can extrapolate a rectangular data set in any direction. All that is required are the regression and penalty matrices, and the appropriate weight matrix. The extrapolation is then effected by (4).

References

Durban, M., Currie, I., and Eilers, P. (2002). Using *P*-splines to smooth two-dimensional Poisson data. *Proceedings 17th IWSM*. Chania, Crete. pp. 207–214.

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

Wand, M. P. (2003). Smoothing and mixed models. *Computational Statistics*. (in press).

One-Sided Tests in Univariate Elliptical Linear Regression Models

Francisco José de A. Cysneiros[1] and Gilberto A. Paula[2]

[1] Departamento de Estatística, Universidade Federal de Pernambuco - Caixa Postal 50749-540, Recife - PE - Brazil, Email: cysneiros@de.ufpe.br
[2] Instituto de Matemática e Estatística, USP - Caixa Postal 66281 (Ag. Cidade de São Paulo), 05311-970 São Paulo - SP - Brazil, Email: giapaula@ime.usp.br

Abstract: We discuss in this paper the problem of testing equality and inequality constraints in univariate elliptical linear regression models. First, the problem of testing the linear equality hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ against the linear inequality hypothesis $H_1 : \mathbf{C}\boldsymbol{\beta} \geq \mathbf{d}$, with at least one strict inequality in H_1 (case 1) and then, $H_1 : \mathbf{C}\boldsymbol{\beta} \geq \mathbf{d}$ against $H_2 : \boldsymbol{\beta} \in I\!R^p - H_1$ (case 2). This class of models includes all symmetric continuous distributions, such as normal, Student-t, Pearson VII, exponential power and logistic, among others. It is commonly used for the analysis of data containing influential or outlying observations with responses supposedly normal. Iterative processes for evaluating the parameters under equality and inequality constraints are presented. Under regular conditions the expressions of the statistics for three asymptotically equivalent statistical tests as well as their asymptotic null distribution are given. An illustrative example with presence of influential observations on the decisions from the statistical tests of different elliptical models is presented. The robustness aspects of such models are discussed.

Keywords: Hypothesis testing; Symmetric distributions; Multivariate one-sided tests; Restricted estimation; Robustness.

1 Univariate Elliptical Linear Models

Let ϵ_i, $i = 1, \ldots, n$, be independent random variables with density function of the form

$$f_{\epsilon_i}(\epsilon) = \frac{1}{\sqrt{\phi}} g\{(\epsilon/\sqrt{\phi})^2\},\ \epsilon \in I\!R, \tag{1}$$

where $\phi > 0$ is the scale parameter, $g : I\!R \to [0, \infty]$ is such that $\int_0^\infty g(u^2)du < \infty$. We shall denote $\epsilon_i \sim El(0, \phi)$. The function $g(.)$ is called density generator (see, for example, Fang, Kotz and Ng, 1990). Consider the linear regression model

$$y_i = \mu_i + \epsilon_i,\ i = 1, \ldots, n, \tag{2}$$

where $\mu_i = \mathbf{x}_i^T\boldsymbol{\beta}$, $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{in})^T$ contains values of p explanatory variables, $y_1, \ldots, y_n$ are the observed response values, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$

is the parameter vector. The model defined by (1)-(2) is called univariate elliptical linear regression model. A joint iterative process to find the unrestricted estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\phi}$ is given by

$$\boldsymbol{\beta}^{(r+1)} = \{\mathbf{X}^T\mathbf{D}(\mathbf{v}^{(r)})\mathbf{X}\}^{-1}\mathbf{X}^T\mathbf{D}(\mathbf{v}^{(r)})\mathbf{y} \qquad \text{and} \tag{3}$$

$$\phi^{(r+1)} = \frac{1}{n}Q_v(\boldsymbol{\beta}^{(r+1)}), \text{ for } r = 0, 1, \ldots, \tag{4}$$

where $Q_v(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{D}(\mathbf{v})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, $\mathbf{D}(\mathbf{v}) = \text{diag}\{v_1, \ldots, v_n\}$, $v_i = -2W_g(u_i)$, $u_i = (y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2/\phi$ and $W_g(u) = g'(u)/g(u)$ with $g(u) = \partial g(u)/\partial u$. We should start the iterative process (3)-(4) with initial values $\boldsymbol{\beta}^{(0)}$ and $\phi^{(0)}$.

2 Restricted estimation

2.1 Equality Constraints

Suppose first we are interested in estimating the parameter vector $\boldsymbol{\beta}$ under k linearly independent restrictions $\mathbf{C}_j^T\boldsymbol{\beta} - d_j = 0$, where $\mathbf{C}_j$, $j = 1, \ldots, k$, are $p \times 1$ vectors and d_j, $j = 1, \ldots, k$, are scalars, both known and fixed. The problem here is to maximize the log-likelihood function $\text{L}(\boldsymbol{\theta})$ subject to the linear constraints $\mathbf{C}\boldsymbol{\beta} - \mathbf{d} = \mathbf{0}$, where $\mathbf{C} = (\mathbf{C}_1^T, \ldots, \mathbf{C}_k^T)^T$ and $\mathbf{d} = (d_1, \ldots, d_k)^T$. Similarly to Nyquist (1991), that investigated this kind of problem in generalized linear models, we shall apply the methodology of penalty functions by considering a quadratic penalty function. The resulting iterative process is given by

$$\begin{aligned}\boldsymbol{\beta}^{0(r+1)} \;=\; & \{\mathbf{X}^T\mathbf{D}(\mathbf{v}^{(r)})\mathbf{X}\}^{-1}\mathbf{X}^T\mathbf{D}(\mathbf{v}^{(r)})\mathbf{y} + \{\mathbf{X}^T\mathbf{D}(\mathbf{v}^{(r)})\mathbf{X}\}^{-1}\mathbf{C}^T \times \\ & \left[\mathbf{C}\{\mathbf{X}^T\mathbf{D}(\mathbf{v}^{(r)})\mathbf{X}\}^{-1}\mathbf{C}^T\right]^{-1} \times \\ & \left[\mathbf{d} - \mathbf{C}\{\mathbf{X}^T\mathbf{D}(\mathbf{v}^{(r)})\mathbf{X}\}^{-1}\mathbf{X}^T\mathbf{D}(\mathbf{v}^{(r)})\mathbf{y}\right],\end{aligned} \tag{5}$$

for $r = 0, 1, \ldots$, where $\phi^{(r+1)}$ is obtained from (4). The authors have developed a library in S-Plus and R to fit univariate elliptical linear models based in some distributions and the iterative process (3-5) and more, some diagnostic graphics. This library is available in the web page `www.de.ufpe.br/~cysneiros/elliptical/elliptical.html`.

2.2 Inequality Constraints

The problem of maximizing log-likelihood functions restricted to linear inequality parameter constraints $\mathbf{C}\boldsymbol{\beta} - \mathbf{d} \geq \mathbf{0}$ have been investigated by various authors (see, for instance, Robertson, Wright and Dykstra, 1988 and Fahrmeir and Klinger, 1994). Our primary interest is to obtain the

maximum likelihood estimate of $\boldsymbol{\beta}$, denoted by $\tilde{\boldsymbol{\beta}}$, in model (1) subject to the constraints $\mathbf{C}\boldsymbol{\beta} - \mathbf{d} \geq \mathbf{0}$; that is, we want to solve the problem $\max_{\{C\beta - d \geq 0\}} \mathrm{L}(\boldsymbol{\beta}, \phi)$. We can apply the Kuhn-Tucker conditions to attain the restricted maximum. These conditions are equivalent to finding $\tilde{\boldsymbol{\beta}}$ from a searching procedure which consists in maximizing $\mathrm{L}(\boldsymbol{\beta}, \phi)$ subject to $\mathbf{C}_j^T\boldsymbol{\beta} - d_j = 0$, $j \in I$, for each $I \subseteq \{1, \ldots, k\}$. The inequality-restricted problem reduces to a equality-restricted problem that may be solved by the procedures given in Section 2.1.

3 One-sided Tests

3.1 Case 1

We shall consider in this section the problem of testing the hypotheses $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ against $H_1 : \mathbf{C}\boldsymbol{\beta} \geq \mathbf{d}$, with at least one strict inequality in H_1. The usual statistics likelihood ratio, Wald and score take, in this case, the forms

$$\begin{aligned}
\xi_{LR} &= 2\left[\frac{n}{2}\log\left(\frac{\hat{\phi}_0}{\tilde{\phi}}\right) + \sum_{i=1}^{n}\log\left\{\frac{g\{(y_i - \mathbf{x}_i^T\tilde{\boldsymbol{\beta}})^2/\tilde{\phi}\}}{g\{(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}^0)^2/\hat{\phi}_0\}}\right\}\right], \\
\xi_W &= \frac{4d_g}{\tilde{\phi}}(\mathbf{C}\tilde{\boldsymbol{\beta}} - \mathbf{d})^T\{\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\}^{-1}(\mathbf{C}\tilde{\boldsymbol{\beta}} - \mathbf{d}) \text{ and} \\
\xi_{SR} &= \frac{\hat{\phi}_0}{4d_g}\{\mathbf{U}_\beta(\hat{\boldsymbol{\beta}}^0, \hat{\phi}_0) - \mathbf{U}_\beta(\tilde{\boldsymbol{\beta}}, \tilde{\phi})\}^T(\mathbf{X}^T\mathbf{X})^{-1}\{\mathbf{U}_\beta(\hat{\boldsymbol{\beta}}^0, \hat{\phi}_0) - \\
&\qquad \mathbf{U}_\beta(\tilde{\boldsymbol{\beta}}, \tilde{\phi})\},
\end{aligned}$$

respectively, where $d_g = E\{W_g^2(Z^2)Z^2\}$ with $Z \sim El(0,1)$ and $\mathbf{U}_\beta(\boldsymbol{\beta}, \phi) = \frac{1}{\phi}\mathbf{X}^T\mathbf{D}(\mathbf{v})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. In addition, suppose the parameter space of $\boldsymbol{\beta}$ is open. Under the regular condition given in Gourieroux and Montfort (1995, Section 21.3) it follows that the statistics ξ_{LR}, ξ_W and ξ_{SR} are asymptotically equivalent as a mixture of chi-square distributions, namely

$$Pr\{\xi_{LR} \geq c\} = \sum_{\ell=0}^{k}\omega(k, \ell; \boldsymbol{\Delta})Pr\{\chi_\ell^2 \geq c\} + o(1), \tag{6}$$

where $c \geq 0$, $\boldsymbol{\Delta} = \mathbf{C}\mathbf{K}_{\beta\beta}^{-1}\mathbf{C}^T$, $\mathbf{K}_{\beta\beta} = \frac{4d_g}{\phi}(\mathbf{X}^T\mathbf{X})$, χ_0^2 denotes the degenerate distribution at the origin and $\omega(k, \ell; \boldsymbol{\Delta})$'s are known as level probabilities which are expressed as functions of correlation coefficients associated with the matrix $\boldsymbol{\Delta}$. These correlation coefficients are the minimum information necessary to compute the asymptotic null distribution given in (6) because $\omega(k, \ell; \boldsymbol{\Delta})$ depends on $\boldsymbol{\Delta}$ only through its correlation matrix. Examining the expression of $\mathbf{K}_{\beta\beta}$ we can conclude that $\omega(k, \ell; \boldsymbol{\Delta})$ does not depend on $\boldsymbol{\beta}$. Then, the distribution given in (6) is unique and consequently invariant in

the elliptical class. This property rarely occurs in other classes of regression models such as generalized linear models (see, for instance, Paula and Sen, 1995).

3.2 Case 2

Now, we shall consider the hypotheses $H_1 : \mathbf{C}\boldsymbol{\beta} \geq \mathbf{d}$ against $H_2 : \boldsymbol{\beta} \in I\!R^p - H_1$. In this case, the usual statistics likelihood ratio, Wald and score take the forms

$$\begin{aligned}
\xi^c_{LR} &= 2\left[\frac{n}{2}\log\left(\frac{\tilde{\phi}}{\hat{\phi}}\right) + \sum_{i=1}^{n}\log\left\{\frac{g\{(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2/\hat{\phi}\}}{g\{(y_i - \mathbf{x}_i^T\tilde{\boldsymbol{\beta}})^2/\tilde{\phi}\}}\right\}\right], \\
\xi^c_{W} &= \frac{4d_g}{\hat{\phi}}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{C}\tilde{\boldsymbol{\beta}})^T\{\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\}^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{C}\tilde{\boldsymbol{\beta}}) \text{ and} \\
\xi^c_{SR} &= \frac{\tilde{\phi}}{4d_g}\mathbf{U}_\beta(\tilde{\boldsymbol{\beta}}, \tilde{\phi})^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{U}_\beta(\tilde{\boldsymbol{\beta}}, \tilde{\phi})^T.
\end{aligned}$$

An important asymptotic result observed in the last section is the lack of functional dependence of $\boldsymbol{\Delta} = \mathbf{C}\mathbf{K}^{-1}_{\beta\beta}\mathbf{C}^T$ on $\boldsymbol{\beta}$. The main consequence of this fact is that the asymptotic null distribution of ξ^c_{LR}, ξ^c_W and ξ^c_{SR} for the purpose of testing H_1 against H_2, is uniquely determined and given by

$$Pr\{\xi^c_{LR} \geq c\} = \sum_{\ell=0}^{k}\omega(k, k-\ell; \boldsymbol{\Delta})Pr\{\chi^2_\ell \geq c\} + o(1). \tag{7}$$

4 Example

We shall reanalyze in this section the example discussed by Ramanathan (1993) on a study in which seven variables were observed in 40 metropolitan areas. The main interest is on regressing the number (in thousands) of subscribers with cable TV (Y) against the number (in thousands) of homes in the area (X_1), the per capita income for each television market with cable (X_2), the installation fee (X_3), the monthly service charge (X_4), the number of television signals carried by each cable system (X_5) and the number of television signals received with good quality without cable (X_6). Because Y corresponds to count data we shall use a square root transformation in order to stabilize the variance of Y. Then, we shall propose the model

$$\sqrt{y_i} = \beta_0 + \sum_{j=1}^{6}\beta_j x_{ji} + \epsilon_i, \qquad i = 1, \ldots, 40,$$

where $\epsilon_i \sim El(0, \phi)$ are mutually independent errors. In addition, it is reasonable to assume some constraints. For example, it is expected that the

number of subscribers decreases as the monthly service charge increases, which leads to the restriction $\beta_4 \leq 0$. Following the same idea for the remaining variables one has the constraints $\beta_1 \geq 0, \beta_2 \geq 0, \beta_3 \leq 0, \beta_5 \geq 0$ and $\beta_6 \leq 0$. Applying one-sided t tests we can notice indications that the coefficients β_2, β_3 and β_4 seem to be individually equal to zero, at the significance level of 5%, while some doubt appears for the coefficient β_5 whose p-value is about 3%. The remaining coefficients β_1 and β_6 are highly significant in the direction of the constraints. Thus, in order to assess if the four coefficients β_2, β_3, β_4 and β_5 are jointly equal to zero, we apply the statistical tests defined in Sections 3.1 to assess the hypotheses $H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ against $H_1 : \beta_2 \geq 0, \beta_3 \leq 0, \beta_4 \leq 0$ and $\beta_5 \geq 0$, with at least one strict inequality in H_1. Our main conclusion of this example based on diagnostic methods is that the transformation $\sqrt{Y}$ seems to stabilize the variance of the responses, but the Student-t with 6 degrees of freedom, exponential power and logistic-II models are less influenced by the outlying observation 14 than the normal model. The one-sided tests based on these three fitted models indicate for the rejection of the null hypothesis at the significance level of 5% while under the normal model the rejection of the null hypothesis becomes evident only after dropping the outlying observation 14. However, the Student-t model seems to be more robust against the influential observation 1 than the other three models. Continuating the selection procedure the Student-t model appears as the best fitted model.

Acknowledgments: The first author received financial support from CAPES and the second author was supported by FAPESP and CNPq, Brazil.

References

Fahrmeir, L. and Klinger, J. (1994). Estimating and testing generalized linear models under inequality restrictions. *Statistical Papers*, **35**, 211-229.

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

Gourieroux, G. and Montfort, A. (1995). *Statistics and Econometric*. Volumes 1 and 2. Cambridge: Cambridge University Press.

Nyquist, H. (1991). Restricted estimation of generalized linear models. *Applied Statistics*, **40**, 133-141

Paula, G. A. and Sen, P. K. (1995). One-sided tests in generalized linear models with parallel regression lines. *Biometrics*, **51**, 1494-1501.

Ramanathan, R. (1993). *Statistical Methods in Econometrics*. New York: Wiley.

Robertson, T., Wright, F.T., and Dykstra, R.L. (1988). *Order Restricted Statistical Inference*. New York: Wiley.

Modeling the Volatility of Assets Returns by GIG Distributions

Joan del Castillo[1] and Anna López[1]

[1] Departament de Matemàtiques, Universitat Autònoma de Barcelona, Campus de la UAB, 08193 Bellaterra, Cerdanyola del Vallés (Barcelona), Spain.

Abstract: In this paper we consider historical volatilities at several time scales measured by different ways, from standard deviations to ranges. Then we modelling it by Generalized Inverse Gaussian distribution (GIG)and GIG mixtures of Gamma distributions. If the basic model setting applies for an asset, then we have to observe a Gamma distribution for sample variance on its returns. Gamma distribution are included in GIG and testing Gamma is a way to check departures from basic hypothesis.

Keywords: Likelihood ratio test; Boundary parameters; Exponential models; Saddlepoint approximation.

1 The Problem

The volatility is one of the most fundamental concepts in finance. It has a major role in risk management and in pricing derivatives. However, it is difficult to obtain a satisfactory estimation for this quantity . The problem arises because the basic model for assets returns assumes constant volatilities and the empirical evidence shows its evolution on time. Several ways are usually considered to measure volatilities: historical volatilities, Black and Scholes implied volatilities, stochastic volatility models, GARCH models... In this paper we consider historical volatilities at several time scales measured by different ways, from standard deviations to ranges. Then we try to modelling it by Generalized Inverse Gaussian distribution (GIG). If the basic model setting applies for an asset, then we have to observe a Gamma distribution for sample variance on its returns. Gamma distribution are included in GIG and testing Gamma is a way to check departures from basic hypothesis.
GIG is a family of infinitely divisible and self-discomposable distributions, hence it is an appropriate model for sums of positive independent and identically distributed random variables and it is compatible with autoregresive models. GIG provides useful models for volatilities, in the context of Lévy processes, and it include many submodels of practical interest: The Gamma model for volatilities, introduced in Madan and Seneta (1990); the positive hyperbola distributions, introduced by Eberlein Kelly (1995)

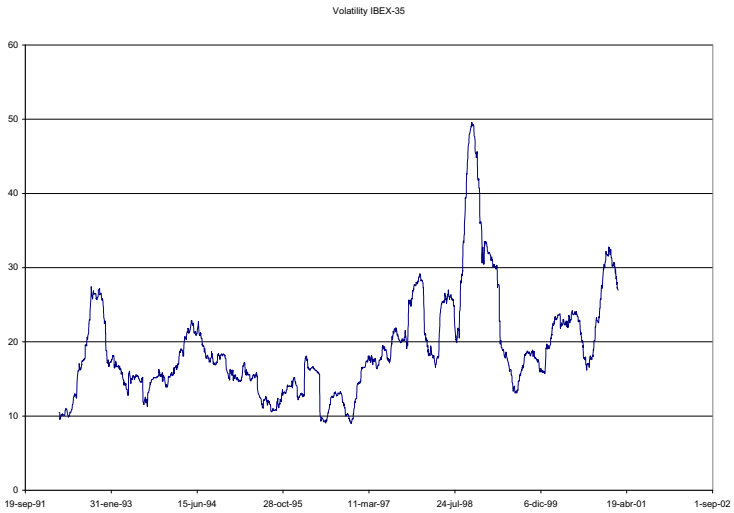


FIGURE 1. *60-day historical volatility estimates*

and the normal inverse Gaussian, introduced by Barndorff-Nielsen (1997). Another relevant submodel of GIG is the reciprocal Gamma family with heavy tailed distributions. All these models are submodels of GIG an can be tested by likelihood ratio test, but the Gamma and reciprocal Gamma distributions appear at the boundary of the family and we face the problem of testing hypothesis in non-regular exponential models.

2 Data Set

The Spanish IBEX-35 index is a value-weighted index comprising the 35 most liquid Spanish stocks traded in the continuous auction marked system. The official derivative marked for risky assets, which is known as MEFF, trades futures contract on the IBEX-35. Trading in derivative marked started in 1992.
For this paper, our data basis is comprised of a time series of daily observed high, low, open and close prices for the IBEX-35 index during the period January 14, 1992 through February 28, 2001.
Figure 1 shows the rolling '60-day historical volatility estimates' that appear to indicate that volatility is changing in some persistent manner over time.

3 The Model: GIG

A generalized inverse Gaussian random variable, $x \sim GIG(\lambda, \chi, \psi)$, has probability density function

$$f(x;\lambda,\chi,\psi) = \frac{(\psi/\chi)^{\lambda/2}}{2K_\lambda(\sqrt{\chi\psi})} x^{\lambda-1} e^{-\frac{1}{2}(\frac{\chi}{x}+\psi x)}, \qquad x > 0 \tag{1}$$

where K_λ is the modified Bessel function of the third kind with index λ. A complete study of the GIG distribution is given in Jörgensen (1982).
A special case is the inverse Gaussian distribution ($\lambda = -\frac{1}{2}$) which arises as the distribution of the first passage time in a Brownian motion with positive drift. The GIG is an exponential model with boundary parameters. On the boundary appears the Gamma distribution ($\chi = 0$, $\lambda > 0$) and the reciprocal Gamma ($\psi = 0$, $\lambda < 0$). For testing Gamma against GIG the parameter of interest χ belongs on the boundary under the null hypothesis. The same situation occurs to test Reciprocal Gamma against GIG. Hence we only consider $\lambda > 0$ since $x \sim GIG(\lambda, \chi, \psi)$ if and only if $x^{-1} \sim GIG(-\lambda, \chi, \psi)$.

4 Main Results

From empirical point of view, it is easily rejected that the estimated volatility follows a Gamma distribution. When GIG distributions are used as a model for volatilities the estimations suggest heavy tailed distributions, as reciprocal Gamma. Moreover, we find a high degree of correlation between return-based and range-based volatility estimates. Then in many places ranges, that are more available in daily financial data, can be used to compare volatilities.
Assuming the GIG model for volatilities and assuming they are locally constant, the distribution of the corresponding return-based estimates is a mixture of Gamma distributions with probability density function

$$f_{S_m^2}(x) = \frac{\left(\frac{m}{2}\right)^{\frac{m}{2}} \psi^{-\frac{m}{4}} \chi^{\frac{\lambda}{2}}}{\Gamma\left(\frac{m}{2}\right) K_\lambda\left(\sqrt{\chi\psi}\right)} \frac{x^{\frac{m}{2}-1}}{(\chi+mx)^{-\frac{\lambda}{2}+\frac{m}{4}}} K_{\lambda-1-\frac{m}{2}}\left(\sqrt{\psi(\chi+mx)}\right). \tag{2}$$

Specially relevant results are obtained on high order properties on testing hypothesis in non regular exponential models that have boundary parameters. The standard regularity conditions do not always work. This fact is closely related to steepness and to the existence of moments of the limit distribution. Testing Gamma or reciprocal Gamma against GIG are examples of this situation. Other examples of interest are conjugate families of non-negative random variables without moments generating function, testing exponentiality against singly truncated normal distribution (Castillo

and Puig, 1999-a) and several situations in reliability theory and survival analysis, see Castillo and Puig (1999-b).
We show that the results of Jensen (1992), on high order properties for likelihood ratio test, can be extended together with the results of Self and Liang (1987) on asymptotic properties for likelihood ratio test under non-standard conditions.

5 Conclusions

1. The constant volatility model for asset returns is clearly regected from empirical evidence.
2. By using a GIG model for volatilities, heavy tailed distributions are suggested.
3. The distribution of sample variance for a GIG mixture of locally constant volatility models is obtained.
4. We prove that saddlepoint methods can be used to improve the likelihood ratio test with boundary parameters.
5. Through simulation we see that the saddlepoint approximation works successfully for small samples for testing Gamma against GIG. Moreover, the simulation results show that the saddlepoint approximation works very well with nuisance parameter too.

6 References

Barndorff-Nielsen, O. (1997). Normal inverse Gaussian distributions and stochastic volatility modelling. *Scandinavian Journal of Statistics*, **24**, 1–13.

Castillo, J. and Puig, P. (1999-a). The Best Test of Exponential Against Singly Truncated Normal Alternatives,*Journal of the American Statistical Association*, **94**, 529–532.

Castillo, J. and Puig, P. (1999-b). Invariant exponential models applied to reliability theory and survival analysis,*Journal of the American Statistical Association*, **94**, 522–528.

Eberlein and Keller (1995). Hyperbolic distributions in finance. *Bernoulli*, **3**, 281–299.

Jensen, J. L. (1992). The modified signed likelihood statistic and saddlepoint approximations. *Biometrika*, **79**(4), 693–703.

Jorgensen, B. (1982). *Statistical Properties of the Generalized Inverse Gaussian Distribution, Lecture Notes in Statistics 9*. New York: Springer-Verlag.

Madan, D.B. and Seneta, E. (1990). The VG model for share market returns. *Journal of Business 1990*, **58**(4), 511–524.

Self, S.G. and Liang, K.Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions, *Journal of the American Statistical Association*, **82**, 605–610.

The Information Matrix Test with Bootstrap-Based Covariance Matrix Estimation

Geert Dhaene[1] and Dirk Hoorelbeke[1]

[1] K.U.Leuven, Department of Economics, Naamsestraat 69, B-3000 Leuven, Belgium. Tel. +32 16 326798. Fax +32 16 326796. Email: geert.dhaene@econ.kuleuven.ac.be.

Abstract: We propose an information matrix test in which the covariance matrix of the vector of indicators is estimated using the parametric bootstrap. Monte Carlo results and theoretical arguments show that its small sample performance is comparable with that of the efficient score form.

Keywords: Information matrix test; Parametric bootstrap.

1 Introduction

The Information Matrix (IM) test, introduced by White (1982), offers a conceptually appealing way to perform omnibus specification testing in parametric models estimated by maximum likelihood. The null hypothesis of the IM test is the information matrix equality, i.e. the hypothesis that the sum of the mathematical expectations of the Hessian matrix of the log-likelihood and the outer product of the gradient vector of the log-likelihood equals zero. An important advantage of the (full) IM test is the fact that it is an omnibus test for misspecification (not requiring the specification of an alternative), e.g. the IM test for the normal regression model tests for heteroscedasticity, skewness and kurtosis.
While the IM test is well known as a general test for misspecification of a parametric likelihood function, its use in applied econometric research is still limited. A major drawback of the IM test is that the asymptotic χ^2 distribution is a very poor approximation to the finite sample distribution of the test statistic. This seriously limits its usefulness in practice. Large deviations from the asymptotic distribution are typical even in relatively large samples, as evidenced by the Monte Carlo experiments reported in Taylor (1987), Orme (1990), Chesher and Spady (1991), Davidson and MacKinnon (1992, 1998), and Horowitz (1994). Several approaches have been proposed to overcome this problem. Chesher and Spady (1991) derive, for specific models, critical values for the IM test statistic that are based on

a higher order Edgeworth expansion. Davidson and MacKinnon (1992) propose a variant of the IM test based on double-length artificial regressions. Their method, however, cannot be applied to models for discrete, censored, or truncated data. Horowitz (1994) proposes bootstrap-based critical values for the IM test. Despite these efforts, computing the correct critical value of an IM test statistic for an arbitrary model is still not particularly easy.

2 Estimating the Covariance Matrix of the IM Test

All existing versions of the IM test rely on some estimate of the asymptotic covariance matrix of the vector of indicators, the differences arising essentially from replacing expectations with sample averages in different parts of the formula for the asymptotic covariance matrix (Orme 1990). Available Monte Carlo evidence shows that the ensuing test statistics have finite sample distributions that are poorly approximated by the χ^2 distribution. Four sources of possible error may be involved in the approximation:

(i) the finite sample distribution of the indicator vector may be non-normal;

(ii) the finite sample covariance matrix of the indicator vector may differ from its asymptotic covariance matrix;

(iii) the unknown parameter is replaced by a consistent estimate in the formula of the asymptotic covariance matrix;

(iv) sample averages replace expectations in parts of the formula for the asymptotic covariance matrix.

In most circumstances, the error sources (i)-(iii) effectively apply to the IM tests. Moreover, the efficient score form is the only one not vulnerable to (iv).

Rather than relying on an asymptotic covariance matrix formula, one may choose to estimate the finite sample covariance matrix of the indicator vector. Although it is simple enough to write the finite sample covariance matrix as an integral, working out the integral analytically is bound to be impossible in all but the simplest models. A simple and feasible alternative is to estimate it by the parametric bootstrap. It is shown that the test statistic with bootstrap covariance matrix has, if the model is correctly specified, an asymptotic (Hotelling's) T^2 distribution with q (the dimension of the indicator vector) and $B-1$ (where B is the number of simulations used to approximate the finite sample distribution of the indicator vector) degrees of freedom. With finite B, the finite sample covariance matrix is estimated with some noise, but the T^2 critical values correct for this. Using this IM test statistic and T^2 critical values, (ii) is eliminated as a source of approximation error.

We have two final remarks. First, the only computational requirement to obtain the statistic with bootstrap covariance matrix is that observations can be generated from the specified density and that the vector of indicators can be computed. The latter can often be extracted without effort from econometric software packages, either as the difference between two information matrix estimates, or as the difference between the inverses of two estimates of the covariance matrix of the MLE. Thus, no analytical work is required before the test can be applied. Second, although Monte Carlo results show that the ERP (error in rejection probability: the difference between the actual and nominal (chosen) rejection probability under the null hypothesis) of the newly proposed test is moderate, it may be advisable in situations with few observations to use bootstrap-based critical values, as suggested by Horowitz (1994) in the context of the IM test. Although this requires a nested bootstrap – the inner bootstrap serves to calculate the covariance matrix estimate – this is nowadays quite feasible: 50 inner and 99 outer bootstrap replications will often suffice (e.g. in the regression model with a constant and three regressors and a sample size of 100, this takes more or less 2 seconds on a P4 2.00GHz using a Matlab program).

3 Monte Carlo Results

We report comparative Monte Carlo results on the finite sample properties of the new statistic, White's (1982), Chesher (1983) and Lancaster's (1984), Orme's (1990) test statistic and the efficient score form. We study the ERP under the null of correct specification as well as the power against a heteroskedastic alternative, both in the linear model and in the probit model. The ERP is displayed using p-value plots (Davidson and MacKinnon, 1998). In order to correct power for ERP, we plot power as a function of actual RP under the null (Davidson and MacKinnon, 1998). In both the linear model and the probit model we find the statistic with bootstrap covariance matrix to have smaller ERP than the other tests. The powers of the efficient score form and the statistic with bootstrap covariance matrix are extremely close to each other, and well above the power of the other included test statistics.

Acknowledgments: Financial support from the Flemish Fund for Scientific Research (grant G.0366.01) is gratefully acknowledged.

References

Chesher, A. (1983). The information matrix test: Simplified calculation via a score test interpretation. *Economics Letters*, **13**, 45–48.

Chesher, A. and Spady, R. (1991). Asymptotic expansions of the information matrix test statistic. *Econometrica*, **59**, 787–815.

Davidson, R. and MacKinnon, J.G. (1992). A new form of the information matrix test. *Econometrica*, **60**, 145–157.

Davidson, R. and MacKinnon, J.G. (1998). Graphical methods for investigating the size and power of hypothesis tests. *The Manchester School*, **66**, 1–26.

Hall, A. (1987). The information matrix test for the linear model. *The Review of Economic Studies*, **54**, 257–263.

Horowitz, J.L. (1994). Bootstrap-based critical values for the information matrix test. *Journal of Econometrics*, **61**, 395–411.

Lancaster, T. (1984). The covariance matrix of the information matrix test. *Econometrica*, **52**, 1051–1053.

Orme, C. (1990). The small-sample performance of the information-matrix test. *Journal of Econometrics*, **46**, 309–331.

Taylor, L.W. (1987). The size bias of the information matrix test. *Economics Letters*, **24**, 63–67.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–26.

Modelling Repeated Paired Comparisons: An Example from the British Household Panel Study

Regina Dittrich[1], Brian Francis[2], and Walter Katzenbeisser[1]

[1] Institute for Statistics, University of Economics, Augasse 2-6, Vienna A-1090, Austria. Email: Regina.Dittrich@wu-wien.ac.at
[2] Centre for Applied Statistics, Lancaster University, Lancaster LA1 4YF, UK

Abstract: This paper introduces a model for repeated paired comparison data. We adopt an approach that converts such data to multiple multivariate responses, which can then be modelled through a log-linear model. Extra parameters can be introduced which can represent e.g. Markovian dependence on the previous time point. Using standard software, we illustrate the technique on attitudinal data from the British Household Panel Survey.

Keywords: Bradley-Terry model; Likert scales; Log-linear model; Temporal dependence; Multinomial responses.

1 Introduction

This paper is concerned with the development of models for repeated paired comparison data, where the same judge compares a set of objects in a paired comparison experiment repeatedly over time. Such data arises naturally in panel surveys, most often in the form of ranked data, where an individual (the judge) might be asked to rank a collection of items or opinions in various sweeps of a survey. Such ranked data can then be converted to a set of paired comparisons by utilising the concept of rank order explosion (Chapman and Staelin, 1982). A second example might be a sports competition such as the Formula 1 motor racing competition, where drivers meet repeatedly in a single year on different race circuits. Here, the "judge" would be the circuit, with the finish order providing the paired comparisons and the repeated yearly visits to each circuit providing the replication. In such data it is natural to wish to consider temporal dependence - the likelihood of a response of a judge at one time point to depend on their response at the previous time point.
Fahrmeir and Tutz (1994) introduced dynamic stochastic models for time-dependent ordered paired comparisons, based on an extension of Kalman filtering and smoothing for dynamic generalized linear models. This has subsequently been extended by Glickman (2001). However the models are complex and time-consuming to fit.

The purpose of this paper is to develop a log-linear approach for (time-dependent) paired comparison data based on the Bradley-Terry model. The approach of this paper is to convert such data to multiple multinomial responses, and echoes the quadratic exponential binary distribution suggested by Cox (1972). This places the model within the Generalized Linear Model (GLM) framework. Parameters representing dependencies can then be added. The advantage of this specification is that model fitting and model checking can easily be done within the GLM framework.

2 Likert Scales and Paired Comparisons

We now consider an associated problem - the analysis of repeated multiple Likert responses over time. We assume that J Likert responses are measured repeatedly over time, and that each Likert response is measured on the same underlying measurement scale. Models for analysing a *single* Likert response and allowing for temporal dependence now exist (Sutradhar and Kovacevic, 2000). Models for multiple ordinal responses at multiple time points (Douglas, 1999) have also been proposed, but the emphasis is on the determination and assessment of a common latent variable rather than the examination of the differences and relative importance of items and such changes over time. Here, we propose an alternative approach. Consider each Likert question to be a separate item. Then for any time point, a set of Likert responses can be converted to a set of paired comparisons simply be examining the response category given to each Likert question. As an example, if the Likert response is greater for item i than for item j, then we determine that item i is preferred to item j. Thus, methods developed for the analysis of repeated paired comparison data may also be appropriate for the analysis of repeated multiple Likert scales.

2.1 An Example

We take as an example a set of social attitude questions from the British Household Panel Study (BHPS). The BHPS is an household-based survey, taking as a base 8,167 selected households in England, Wales and most of Scotland. We concentrated on a question which measures concern about various social and political issues of contemporary relevance. This question has so far been administered three times - in 1992, 1994 and 1996 - and consist of a series of four-point Likert scales which give the absolute level of concern (a great deal, a fair amount, not very much, not at all). We have chosen the following three items, *destruction of the ozone layer*, *rate of unemployment*, and *declining moral standards*. 4,155 panel members responded to all three waves of the survey and gave complete responses.

3 A Log-linear Representation for Time-dependent Paired Comparisons

Consider a paired comparison experiment where J objects $O_1, \ldots, O_J$ have to be compared repeatedly over time points t, $t = 1, 2, \ldots, T$, by N judges, where it is assumed that all judges respond at each time point to all paired comparisons. At each time point t, this experiment results in $\binom{J}{2}$ paired comparisons. We represent the 'worth' of the object O_i at time t on an underlying latent scale by the parameter π_{it}, with $\sum_{i=1}^{J} \pi_{it} = 1$.
For the comparison of objects O_i and O_j at time point t, the response to the experiment is represented by the random variable Y_{ijt}, defined by

$$Y_{ijt} = \begin{cases} -1 & \text{if object } O_j \text{ is preferred over } O_i \text{ at time } t, \\ 0 & \text{if there is no preference between } O_i \text{ and } O_j \text{ at time } t, \\ 1 & \text{if object } O_i \text{ is preferred over } O_j \text{ at time } t. \end{cases}$$

In terms of the random variables Y_{ijt}, the experiment results for a given judge in a response pattern vector of length $T \times \binom{J}{2}$ which can be written in a pre-defined fixed order as

$$\mathbf{y} = (y_{121}, \ldots, y_{12T}; y_{131}, \ldots, y_{13T}; \ldots; y_{J-1\,,J1}, \ldots, y_{J-1\,,JT}) .$$

For any judge, the observed response pattern will be one of the $L = 3^{T\binom{J}{2}}$ possible response pattern vectors $\mathbf{y}_\ell$, $\ell = 1, 2, \ldots, L$, with each element consisting of one the values $\{-1, 0, 1\}$. For example, $\mathbf{y}_1 = (-1, -1, \ldots, -1, -1)$.
For any Y_{ijt} the response is ordinal and the Adjacent Categories model (Böckenholt and Dillon, 1997) is a suitable basis for our model. This model can be written as

$$P\{Y_{ijt} = y_{ijt}\} = \Delta_{ijt} \left(\frac{\pi_{it}}{\pi_{jt}} \right)^{y_{ijt}} (\nu_{0t})^{\,1-|y_{ijt}|}\,,\; y_{ijt} \in \{-1, 0, 1\}\,, \qquad (1)$$

where Δ_{ijt} denotes a normalising constant in order to make the probabilities sum to unity, ν_{0t} can be interpreted as a parameter representing *no decision* at time point t.
We assume that decisions concerning different object pairs are independent. We can therefore write the joint distribution of the Y_{ijt} as

$$P\{\mathbf{Y} = \mathbf{y}\} = \prod_{i<j} P\{Y_{ij1} = y_{ij1}, Y_{ij2} = y_{ij2}, \ldots, Y_{ijT} = y_{ijT}\}\,. \qquad (2)$$

If there is no temporal dependence, this expression can be further factorised into

$$P\{\mathbf{Y} = \mathbf{y}\} = \prod_{i<j} \prod_{t=1}^{T} P\{Y_{ijt} = y_{ijt}\}\,. \qquad (3)$$

Temporal dependence is introduced by assuming that only the previous decision (at time $t-1$) has an influence on the decision at time t for the given comparison of objects i and j. This is represented by additional parameters $\theta_{ij|t-1,t}$ which are introduced as follows:

$$\begin{aligned} & P\{Y_{ij1} = y_{ij1}, \ldots, Y_{ijT} = y_{ijT}\} \\ &= \left[\Delta_{ij}\left(\frac{\pi_{i1}}{\pi_{j1}}\right)^{y_{ij1}} (\nu_{01})^{\,1-|y_{ij1}|}\right] \\ &\quad \times \prod_{t=2}^{T}\left(\frac{\pi_{it}}{\pi_{jt}}\right)^{y_{ijt}} (\nu_{0t})^{\,1-|y_{ijt}|} \exp\{\theta_{ij|t-1,t} y_{ij,t-1} y_{ijt}\}\,, \end{aligned} \tag{4}$$

where $\Delta_{ij} = \prod_t \Delta_{ijt}$.
The joint distribution of $\mathbf{Y}$ can then be obtained by using equation (2). Note that this is the probability of the independence model (3) augmented by a multiplicative term that introduces time dependencies between the Y's.
The probability of any particular response pattern $p(\mathbf{y}_\ell)$ can then be obtained by substituting values of the y_{ijt} into the above expression, and the expected value $m_\ell = N \times p(\mathbf{y}_\ell)$ can be calculated. Taking logs converts this model into a log-linear model and can be estimated using standard software using a log-link and Poisson distribution. The design matrix consists of a column for a nuisance parameter δ and sets of columns for the λ, γ, θ parameters, where $\lambda_{it} = \ln \pi_{it}$, $\gamma_{0t} = \ln \nu_{0t}$. Further details and discussion of other dependencies are given in Dittrich et al (2003).

4 Results

We return to the BHPS data introduced earlier. With three items and three time points, and with three possible outcomes to each paired comparison, we have 19,683 possible responses in the response set. We fit two models - the temporal independence model (deviance 29588 on 19673 df) and the temporal dependence model (deviance 24550 on 19667 df). It is clear from the change in deviance that there is strong temporal dependence. The θ parameters are all close to one with those measuring dependence in the ozone-morals comparison (1.257 and 1.228) exhibiting the highest dependence and thus the most stability in response over time. We can also examine the worths $\pi_{it} = exp(\lambda_{it})/\sum_j exp(\lambda_j t)$ for both the independence and dependence models (Figure 1). Both plots show that concern about the high unemployment rate is decreasing over the period of observation, whereas concern about moral standards declining is increasing over the period. Small changes are noted in the worth parameters in shifting from the independence to the dependence model.

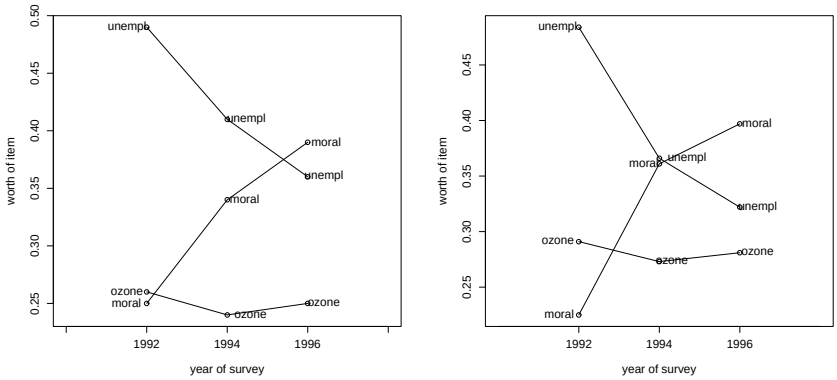


FIGURE 1. *Worth parameters for the models of temporal independence (left) and temporal dependence (right)*

References

Böckenholt, U. and Dillon, W. (1997). Modelling within-subject dependencies in ordinal paired comparison data. *Psychometrika*, **62**, 411–434.

Chapman, R.G. and Staelin, R. (1982). Exploiting rank ordered choice set data within the stochastic utility model. *Marketing Research*, **19**, 288–301.

Cox, D. (1972). The analysis of multivariate binary data. *Applied Statistics.*, **21**, 113–120.

Dittrich, R., Francis, B., and Katzenbeisser, W. (2003). Temporal dependence in longitudinal paired comparisons. Submitted.

Douglas, J.A. (1999). Item response models for longitudinal quality of life data in clinical trials. *Statistics in Medicine*, **18**, 2917–2931.

Fahrmeir, L. and Tutz, G. (1994). Dynamic stochastic models for time-dependent ordered paired comparison systems. *Journal of the American Statistical Association*, **89**, 1438–1449.

Glickman, M.E. (2001). Dynamic paired comparison models with stochastic variances. *Journal of Applied Statistics*, **28**, 673–689.

Sutradhar, B.C. and Kovacevic, M. (2000). Analysing longitudinal survey data: Generalised estimating equations approach. *Biometrika*, **87**, 837-848.

Mixture Models for Background Estimation

Paul H.C. Eilers[1]

[1] Department of Medical Statistics, Leiden University Medical Centre

Abstract: Background estimation is common problem in many types of measurements. Asymmetric smoothing, using either an L_2 or an L_1 is an effective ad-hoc solution. But it has the disadvantage that it only specifies an algorithm, not a statistical model. To remedy this, a mixture model for baseline, noise and (positive) signal is introduced. The EM algorithm is used for estimation. The model works well and its use is illustrated on real data.

Keywords: Asymmetric least squares; P-splines; Quantile smoothing.

1 Introduction

In many types of measurements we encounter an unavoidable background signal. It may be caused by drifting of the instrument over time, like in a chromatogram, it may be a signal contribution due to interfering substances, as in many (optical) spectra, or it may be background fluorescence in a microarray image. In many cases the background can be modelled — in words — as a slowly varying baseline on which the real signal is superposed. A picture shows stretches of "pure" baseline, alternating with the peaks of the signal. In most cases the chemistry or physics of the set-up dictates that the real signal be strictly positive or strictly negative, but the picture may be blurred by noise. If the background is relatively weak, it does little harm, but in many real-life applications good background correction can improve detection limits appreciably. Some instruments give the operator the option to correct the data interactively. The human eye is a wonderful pattern recognition machine; a trained operator can indicates baseline stretches with a mouse. A computer program then connects these with straight lines or splines to construct a complete baseline. Of course, such a semi-manual approach is not the most desirable one. It is even impossible in many high-throughput systems or in completely automatic processes. Thus there is a need for reliable procedures for automatic background estimation.

Simple low-pass filtering will not work, because of the basic asymmetry of the situation. The real (partly high-frequency) signal deviates only in one direction from the low-frequency baseline. A statistical model has to respect this. In this paper I discuss several promising approaches:

- Asymmetric least squares. A weighted least squares smoother or parametric curve fitter is used, but positive and negative residuals get different weights (with a ratio of 100 or more). A simple iterative algorithm works well to fit this model.
- Percentile smoothing. Again the weights of positive and negative residuals are different, but now an L_1 norm (sum of absolute values) is used. A large-scale linear program has to be solved (using the interior point method).
- Mixture modelling. An explicit model for baseline, noise and the signal distribution is set up and fitted by an EM algorithm.

2 Asymmetric Smoothing

To simplify the presentation, I use P-splines (Eilers and Marx, 1996), my favorite smoother, but local likelihood or smoothing splines might be used equally well. I first discuss standard smoothing. Let y be a measured series, with positions x, and let $\mu = B\alpha$, where B is a B-spline basis, computed on x. Minimizing $|y-\mu|^2$ leads to the normal equations $B'B\hat{\alpha} = B'y$. B is chosen to be "rich", i.e. it would generally overfit, giving $\hat{\mu} = B\hat{\alpha}$ that is less smooth than desired. To increase the smoothness of μ, the penalized sum of squares $|y - B\alpha|^2 + \lambda|D_d\alpha|^2$ is minimized. Here D_d is the matrix that forms differences of order d. The explicit solution follows from the modified normal equations $(B'B + \lambda D_d'D_d)\hat{\alpha} = B'y$. With λ we can tune the smoothness of $\hat{\mu} = B\hat{\alpha}$. Of course, μ will go more or less through the "middle" of y.

To get an asymmetric result, we introduce adaptive weights w_i: $w_i = a$ if $y_i > \mu_i$ and $w_i = 1 - a$ if $y_i \leq \mu_i$, with $0 < a < 1$. The goal function is $(y - B\alpha)'W(y - B\alpha) + \lambda|D_a\alpha|^2$, with W a (diagonal) matrix with the asymmetric weights on the diagonal. If the real signal has positive peaks (above the baseline) $a = 0.01$ or $a = 0.001$ is used, but when they are negative (below the baseline) $a = 0.99$ or $a = 0.999$. Experience has shown that with visual inspection a very good baseline fit can be obtained. A simple algorithm works well: given the weights w, finding $\hat{\mu}$ is just a case of weighted (penalized) linear regression. And given $\hat{\mu}$, the computation of the weights is trivial. Starting with all weights equal to 1, the two computations are alternated until convergence. The convexity of the goal function guarantees that this will take place exactly. In practice about 10 iterations are sufficient.

A modification of this scheme uses the L_1 norm in the goal $|W(y - B\alpha)| + \lambda|D_d\alpha|$, where W is a diagonal matrix, with the adaptive weights w_i on its diagonal. We need a linear programming algorithm to solve this problem. The interior point method works well (Eilers, 2000). Notice that essentially we are estimating a smooth low-percentile curve.

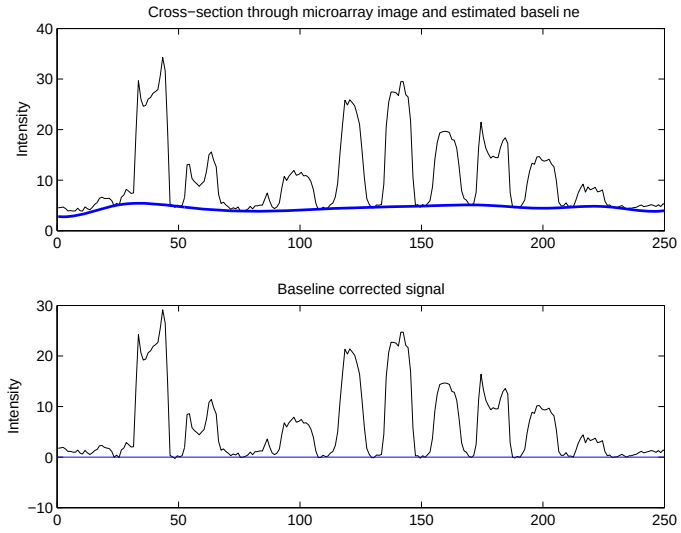


FIGURE 1. *Background estimation with asymmetric least squares smoothing. The upper panel shows the data with a thin line and the estimated baseline with a thick line. The lower panel shows the difference between data and estimated baseline.*

A disadvantage of these algorithms is that the choice of the parameters p and λ is made visually. This is not a great problem in a semi-interactive set-up. Figure 1 shows an example. The data are a cross-section through a part of the fluorescence image of a cDNA microarray. The background was estimated with asymmetric least squares smoothing, using a basis of 13 cubic B-splines, $\lambda = 10^{-4}$, and $a = 0.001$.

3 A Mixture Model

Asymmetric smoothing is defined algorithmically and it is not very clear which criteria to use to optimize the parameters (a for asymmetry and λ for smoothness). This section presents a mixture model with explicit components for baseline, noise and signal.
To simplify the presentation, we first consider the case of a constant but unknown background level μ. We assume normally distributed noise with unknown variance σ^2 and a signal with an unknown distribution $h(.)$, which is only supported on the positive real axis. The mixture model is:

$$f(y) = \pi g(y|\mu, \sigma) + (1 - \pi)h(y - \mu), \tag{1}$$

where π is an unknown mixing ratio and $g(.)$ stands for a normal density.

The EM algorithm is attractive here. Suppose we knew approximations to all parts of (1). Then we could compute the approximate posterior probability p_i that observation y_i comes from density $g(y_i|\mu, \sigma)$ as

$$p_i = \frac{\pi g(y_i|\mu, \sigma)}{\pi g(y_i|\mu, \sigma) + (1 - \pi)h(y_i - \mu)}. \tag{2}$$

Then we can use p and y to compute weighted estimates of μ and σ. And we can feed $1 - p$ to a density smoother to estimate $h(.)$ and take the average of p to estimate π. Hopefully this gives improved approximations, so by repeating these steps we would finally arrive at the solution. Experience with real and simulated data has shown that this is actually the case.
The amount of smoothing for estimating $h(.)$ is important. Oscillations can occur when it is not chosen strong enough. Actually we can simplify this component of the model to an exponential (or even a uniform) distribution without any problem. We have no interest in the actual signal distribution, but only need good estimates of the probabilities p. For y_i near μ, p_i is very near to 1, and at a distance larger than 3σ from μ it will be essentially 0, whatever the shape of $h(.)$. The simplification will have some influence on weights not near 0 or 1, but they are a minority.
Now it is easy to see how more complicated baseline models can be constructed: specify $\mu(x)$ as a polynomial or P-spline model in x. This model and the EM algorithm work remarkably efficient and effective. For the microarray data it looks very similar to Figure 1. Figure 2 shows an example with a rather strongly fluctuating baseline, using 100 cubic B-splines and $\lambda = 0.001$.
To correct the background of an image it is not attractive to apply the algorithm to each column (or row) separately, because small jumps can occur when going from one column (row) to the next. The model allows straightforward extension to a two-dimensional model, using tensor products of P-splines (Durban, Currie and Eilers, 2002). The implementation described there would not work here, because in an intermediate step a regressor matrix of size m by n is formed, with m the number of observations and n the number of basis functions. For an image with 500 by 500 pixels and a 10 by 10 grid of tensor products this leads to a matrix with 25 million elements, taking (too) much space and time. Very recently we (Eilers, Currie and Durban) have developed an extremely fast algorithm for weighted tensor product smoothing of data on a grid that eliminates this intermediary step. Details will be reported elsewhere.

4 Discussion

The mixture model works well. It solves one problem: the choice of a measure of asymmetry, because it follows implicitly from the parameters π, σ and the model for distribution $h(.)$. But we are still left with the penalty

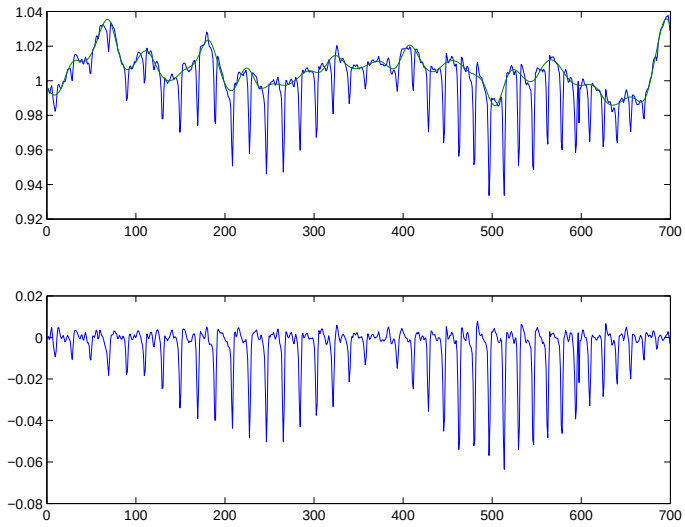


FIGURE 2. *Background estimation with the mixture model. The upper panel shows the data with and the estimated baseline. The lower panel shows the difference between data and estimated baseline.*

parameter λ. Cross-validation seems a natural choice here. We fit the model to a subset of the data, chosen randomly or systematically, like only the odd observations, and compute the log-likelihood of the left-out part of the data. Changing the penalty parameter λ on a grid and computing the cross-validation likelihood for each value will give a curve that hopefully shows a global maximum. Experiments with simulated data seem to indicate that this can work well. But the simulations use independent noise. In real data the noise is frequently correlated, leading to complications. More work is needed here.

References

Durban, M., Currie, I., and Eilers, P.H.C. (2002). Using P-splines to smooth two-dimensional Poisson data. *Proceedings of the 17th International Workshop on Statistical Modelling.*

Eilers, P.H.C. (2000). Robust and Quantile Smoothing with P-splines and the L_1 Norm. *Proceedings of the 15th International Workshop on Statistical Modelling, Barcelona.*.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible Smoothing with Splines and Penalties (with discussion). *Statistical Science*, **11**, 89–121.

A Two-step Estimator for Censored Linear Models with Measurement Errors on Covariates

Anna Espinal [1] and Albert Satorra[2]

[1] Universitat Autonoma de Barcelona. Edifici D, campus Bellaterra. 08193 Cerdanyola del Valles, Spain
[2] Universitat Pompeu Fabra. Ramon Trias Fargas, 25-27. 08005 Barcelona, Spain

Abstract: A typical analysis of survival data assesses the impact of several explanatory variables on a time duration response variable. The standard methodology for such analysis assumes that the explanatory variables, or covariates, are measured without error. We deal with the analysis of data in which a response variable is right-censored and some covariates are contaminated with measurement error. We assume a log-linear model with a right-censored response, and a set of covariates some of them measured with error. To obtain consistent estimates of the regression parameters that takes measurement error into account, we propose a sequential procedure. The performance of the two-step estimator is studied using simulated data. Finally, standard errors are also obtained.

Keywords: Right-censoring; Censored linear model; Measurement errors.

1 Introduction

A frequent problem in statistics is to obtain the estimates of the regression parameters, that is, to assess the effects of a set of covariates on a response variable. In survival analysis, the presence of censoring requires specialized methods for estimating unknown parameters. For linear models, we emphasize the procedures that are modifications of Least Squares (LS) methods to accommodate censored values of the response (see, e.g. Buckley and James 1979). A common assumption underlying these methods is that covariates are measured precisely.

Even though there is a wide range of methodologies for estimating the regression parameters taking into account measurement errors (see Fuller, 1987 or Carrol, Rupert and Stefanski, 1995), all of them are based on the values for the dependent variable when no censoring is present.

We propose a method for estimating censored linear models with measurement errors on covariates based on a combined procedure that merges known results from measurement error theory with methods for censored data. We describe a two-step approach for obtaining consistent estimates of the regression parameters.

2 The Model

We consider a non-negative and continuous random variable T (this is the time elapsed in a certain state) and a set of explanatory variables $\{X_1^\star, \cdots, X_p^\star\}$, also called covariates.
Let $Y = \log T$ be the log-transformation of the true duration T. Consider $y_1, \cdots, y_n$, independent realizations of Y such that y_i is related to the vector of covariates $\mathbf{x_i^\star}$ as

$$y_i = \mathbf{x}_i^{\star\prime}\boldsymbol{\beta} + w_i, \quad i = 1, \cdots, n \tag{1}$$

where $\boldsymbol{\beta}$ is the vector of unknown parameters and $w_1, \cdots, w_n$ are i.i.d. realizations of a disturbance term W with variance σ_w^2 and mean not necessarily zero. We assume that W and $X_j^\star, j = 1, \cdots, p$, are independent random variables.
We assume a right censorship model, that is, our observable duration for the ith individual is

$$z_i = \min\{y_i, c_i\}, \quad i = 1, \cdots, n \tag{2}$$

where $c_1, \cdots, c_n$ are independent realizations of a random variable C (in this case c_i represents the log-transformed censored time for individual i). Here we assume that the censoring mechanism is not informative. The indicator of censoring is given by $\delta_i = \mathbf{1}_{\{y_i \leq c_i\}}, \quad i = 1, \cdots, n$.
The model defined by (1) and (2) stated for analyzing data of the form $\{(z_i, \delta_i, \mathbf{x}_i^{\star\prime}), \quad i = 1, 2, \cdots, n\}$ is usually known as the censored linear model (see, e.g. Breiman, Tsur and Zemel, 1993). From now we refer to it as CLM.
Here we consider a CLM with an additional assumption of possible measurement error in the covariates. Thus we assume that variables $X_j^\star$ may be observed only indirectly, through covariates X_j, $j = 1, \cdots, p$. The relationship between the observed covariates $\mathbf{x}_i$ for the ith individual and the true value of the covariates $\mathbf{x}_i^\star$ is defined by the measurement error model:

$$\mathbf{x}_i = \mathbf{x}_i^\star + \mathbf{u}_i, \quad i = 1, \cdots, n \tag{3}$$

where $\mathbf{u}_1, \cdots, \mathbf{u}_n$ are i.i.d. realizations of the random vector $\mathbf{U} = (U_1, \cdots, U_p)'$ with zero mean and known covariance matrix Σ_{uu}. We also assume that $\mathbf{U}$ is independent of $\mathbf{X}$ and W.

3 The Two-step Estimator

The two-step estimator gives unbiased estimates of the regression coefficients of the model defined by (1), (2) and (3). The method modifies the standard procedures of estimation for linear measurement error models in

order to account for censoring. The first step of the method takes into account the presence of censoring in the data. The second step consists of estimating a linear measurement error model defined by (1) and (3). Once both steps have been performed, unbiased estimators of the regression coefficient in model (1) are obtained.

3.1 Estimated censored values: Step 1

In this step, we ignore the measurement error, in the sense that we state the survival model defined by

$$\begin{aligned} y_i &= \mathbf{x}_i'\boldsymbol{\gamma} + \epsilon_i \\ z_i &= \min\{y_i, c_i\} \\ \delta_i &= \mathbf{1}_{\{y_i \leq c_i\}} \end{aligned} \tag{4}$$

where $\mathbf{x}_i$ is the vector of explanatory variables for individual i (here we are using their observed values) and $\boldsymbol{\gamma}$ are the regression coefficients. Note that we change the notation for the parameters because in the observed model (4) the parameters are not the same as those in the true model defined in (1) and (2).

We note that (4) is a CLM. Thus, to obtain consistent estimates of parameter $\boldsymbol{\gamma}$ we suggest, based on the ease of implementation in practical situations including multiple regression, using the method proposed by Schneider and Weissfeld (1986). Say, $\hat{\boldsymbol{\gamma}}$ the estimator of $\boldsymbol{\gamma}$ obtained by applying this method to the model defined in (4).

In this step we want to deal with the censoring of the response variable. For this reason, we are not interested in the estimator $\hat{\boldsymbol{\gamma}}$ but in linear predictions, conditional on $\mathbf{x}_i$, for z_i based on model (4), say $\hat{z}_i = \mathbf{x}_i'\hat{\boldsymbol{\gamma}},\ i = 1, \cdots, n$. This leads to the following result for the values $(\hat{z}_1, \cdots, \hat{z}_n)$:

Result 1.
The $(\hat{z}_1, \cdots, \hat{z}_n)$ are estimators of the censored response variable such that

$$\hat{\boldsymbol{\kappa}}_{xy} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \hat{z}_i \tag{5}$$

is a consistent estimator of $\mathbf{k}_{xy} = \mathrm{E}(X'Y)$, where Y is the true response variable. That is,

$$\hat{\boldsymbol{\kappa}}_{xy} \xrightarrow{P} \mathbf{k}_{xy}.$$

The usefulness of this estimator $\hat{\boldsymbol{\kappa}}_{xy}$ is based on the following argument. If the response variable in a linear model is censored, for the observed z_i the matrix of the raw mean squares and products $\mathbf{K}_{xz} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i z_i$ is not a consistent estimator of $\mathbf{k}_{xy}$.

3.2 Errors-in-variables Model: Step 2

In this step, we define the estimator of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}$, using methodologies for estimating linear measurement error models (see Fuller, 1987). The proposed procedure is based on the estimator $\hat{\boldsymbol{\kappa}}_{xy}$ defined in step 1.
We consider the errors-in-variables model

$$\begin{aligned} y_i &= \mathbf{x}_i^{\star\prime}\boldsymbol{\beta} + w_i \\ \mathbf{x}_i &= \mathbf{x}_i^{\star} + \mathbf{u}_i. \end{aligned} \tag{6}$$

where the covariance matrix of $\mathbf{U} = (U_1, \cdots, U_p)$, denoted by Σ_{uu}, is known. Then for the standard case where y_i are observed for all $i = 1, \cdots, n$, a consistent estimator of $\boldsymbol{\beta}$ is defined as (see Fuller, 1987)

$$\hat{\boldsymbol{\beta}} = (\mathbf{K}_{xx} - \Sigma_{uu})^{-1}\,\mathbf{K}_{xy} \tag{7}$$

where $\mathbf{K}_{xx} = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'$ and $\mathbf{K}_{xy} = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i y_i$. As before, $\mathbf{K}_{xx}$ and $\mathbf{K}_{xy}$ denote the matrix of the raw mean squares and products.

Result 2.
The proposed estimator of $\boldsymbol{\beta}$ defined as

$$\hat{\boldsymbol{\beta}} = (\mathbf{K}_{xx} - \Sigma_{uu})^{-1}\,\hat{\boldsymbol{\kappa}}_{xy} \tag{8}$$

is a consistent estimator of $\boldsymbol{\beta}$, where $\hat{\boldsymbol{\kappa}}_{xy}$ is the estimator computed in step 1.
The estimator derived by steps 1 and 2 is a consistent estimator of the regression parameters of model (1), (2) and (3). It is called the two-step estimator. The performance of the estimator is explored using simulations.

3.3 Standard Errors

In order to obtain the standard errors of the two-step estimator we assume first uncensored observations only. In such a case, asymptotic robust standard errors may be computed using the normal theory estimates (see Satorra, 1992).
However, in the presence of censoring, the usual formulae for standard errors in linear measurement error models do not apply. We advocate computing standard errors using bootstrap methods.
Table 1 shows 5% and 10% tails of the empirical distribution of the z-statistic of the two-step estimator defined in (8). The results indicate that these empirical values remain close to the theoretical ones when there is censoring in the response and measurement error on covariates.

TABLE 1. *Monte Carlo results with 20% of Type I of censoring.* $B(\cdot)$ *is the bias of the estimator,* $V(\cdot)$ *denotes the estimated variance of the* z*-statistic and 5%−tail, 10%−tail are the empirical* $P(|z| > 1.96)$ *and* $P(|z| > 1.65)$*, respectively. Population value of parameters* $\beta_0 = 3, \beta_1 = 1$.

	$\hat{\beta}_0$				$\hat{\beta}_1$			
k	B($\hat{\beta}_0$)	V(z)	5% *tail*	10% *tail*	B($\hat{\beta}_1$)	V(z)	5% *tail*	10% *tail*
				$n = 100$				
1	–.011	1.04	6.20	11.20	–.011	1.09	6.60	10.60
.8	–.009	.95	4.00	9.20	.000	.92	3.60	8.80
.6	–.004	.79	3.20	5.80	.027	.85	4.20	8.00
.4	–.017	.52	1.40	3.80	.068	.48	1.80	4.20
				$n = 500$				
1	–.003	1.04	5.20	10.40	–.006	1.07	6.20	10.80
.8	–.008	1.09	6.80	11.80	–.006	1.04	6.00	10.60
.6	–.004	.98	5.00	10.00	.002	1.03	5.00	10.00
.4	–.010	.87	5.00	10.00	.005	1.01	4.60	9.60
				$n = 1000$				
1	–.005	1.14	7.40	12.20	–.006	1.02	5.80	12.40
.8	–.007	1.03	6.40	11.80	–.007	1.17	6.40	11.20
.6	–.008	1.06	6.20	11.60	–.009	1.05	7.20	12.20
.4	–.011	1.02	6.20	11.40	–.008	1.01	6.20	11.80

References

Breiman, L., Tsur, Y., and Zemel, A. (1993). On a simple estimation procedure for censored regression models with known error distributions. *Annals of Statistics*, **21** 1711–1720.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika*, **66**, 429-436.

Carrol, R., Rupert, D., and Stefanski, L. (1995). *Measurement Error in Nonlinear Models*. New York: Chapman & Hall.

Fuller, W. (1987). *Measurement Error Models*. New York: Wiley.

Satorra, A. (1992). Asymptotic robust inferences in the analysis of mean and covariance structures. *Sociological Methodology*, **22**, 249–278.

Schneider, H. and Weissfeld, L. (1986). Estimation in linear models with censored data. *Biometrika*, **73**, 741–754.

Hierarchical Modelling Approach for Risk Assessment in Developmental Toxicity Studies.

C. Faes[1], H. Geys[1], M. Aerts[1], and G. Molenberghs[1]

[1] Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium. Email: christel.faes@luc.ac.be

Abstract: Within the past decade, there has been an increasing interest in the problem of joint analysis of clustered multiple outcome data, motivated by developmental toxicity applications (Fitzmaurice and Laird 1995, Gueorguieva and Agresti 2001, Molenberghs and Ryan 1999, Regan and Catalano 1999, Aerts et al 2002). So far however, one has tackled the challenges in this setting only partly each time making different restricting assumptions (e.g. restriction to viable fetuses only). Ideally, a model should take the complete correlated hierarchical structure of the data into account. A hierarchical bayesian method will be discussed in this context. Once a suitable model is selected, it can serve as basis for quantitative risk assessment.

Keywords: Toxicology; Benchmark dose; Hierarchical model.

1 Introduction

Lately, society has become increasingly concerned about problems related to fertility and pregnancy, birth defects and developmental abnormalities. Questions are raised about the potential risk of chemical compounds and other environmental agents on the development of fetuses. Consequently, regulatory agencies such as the U.S. Environmental Protection Agency and the Food and Drug Administration have given increased priority to reproductive and developmental toxicity research, in order to investigate the causes of these problems and to assess the potential adverse effects of exposure on the developing fetuses.
However, because of ethical reasons, reliable epidemiological information of adverse effects on fetal development may often be limited or unavailable. As an alternative, laboratory experiments in small mammalian species can be conducted in advance of human exposure (Williams and Ryan 1996). In developmental toxicity studies with a Segment II design, pregnant animals are exposed during the period of major organogenesis and structural development to a compound of interest. Dose levels for this design typically consist of a control group and three or four exposed groups, each with 20

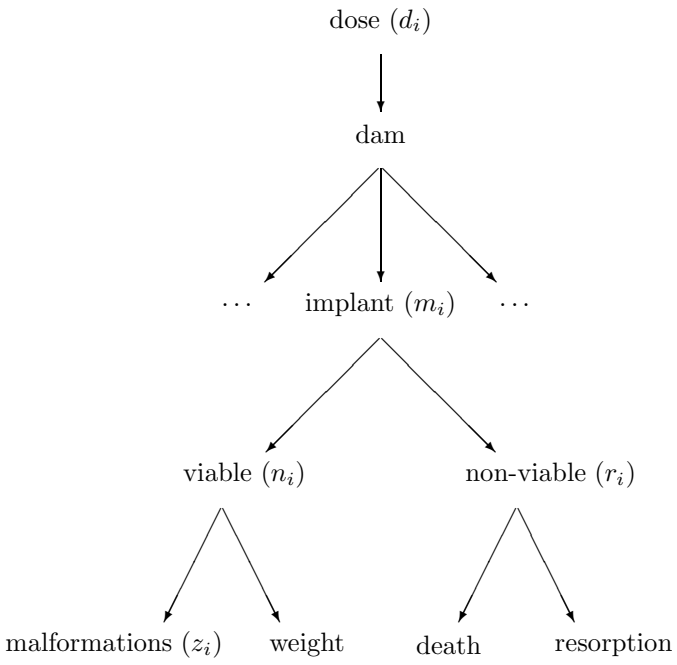


FIGURE 1. *Data Structure of Developmental Toxicity Studies*

to 30 pregnant animals. The dams are sacrificed just prior to normal delivery, at which time the uterus is removed and the contents are thoroughly examined for the occurrence of defects. The viable fetuses are measured for birth weight and examined carefully for the presence of malformation.
The analysis of developmental toxicity data raises a number of challenges (Molenberghs et al 1998). Since deleterious events can occur at several points in development, an interesting aspect lies in the staging or hierarchy of possible adverse fetal outcomes (Williams and Ryan 1996). Figure 1 illustrates the data structure. A toxic insult early in gestation may result in a resorbed fetus. If the implant survives being absorbed, the developing fetus is at risk of fetal death. If the fetus survives the entire gestation period, growth reduction such as low birth weight may occur. The fetus may also exhibit one or more types of malformation. Ultimately, a model should take into account this hierarchical structure. In addition, because of genetic similarity and the same treatment conditions, offsprings of the same mother behave more alike than those of another mother, i.e., the litter effect. Thus, responses on different fetuses within a cluster are likely to be correlated.

2 Risk Assessment

The primary goal of these studies is to determine a safe level of exposure. Recent techniques for risk assessment in this area are based on fitting dose-response models and estimating the dose corresponding to a certain increase in risk of an adverse effect over background, i.e. benchmark dose (Crump 1984).
In case of multiple outcomes, the outcomes are often examined individually, using appropriate methods to account for the correlation, and regulation of exposure is based on the most sensitive outcome. It has been found, however, that a clear pattern of correlation exists between all the outcomes (Ryan et al. 1991), so that risk assessment based on a joint model may be more appropriate. The model must both incorporate the correlation between the outcomes, as well as the correlation due to clustering. Estimation of the risk, will be illustrated in Section 4.

3 Modelling Approach

Until now, most models have looked only to a small part of the hierarchical structure, and assumed that the response distribution for the malformation outcomes and weight outcomes is independent of the cluster size. The analysis of developmental toxicity data has usually been conducted on the number of viable fetuses only. In other models, the litter-size was included as a covariate in modelling these response probabilities (Williams 1987, Rai and Van Ryzin 1985, Catalano and Ryan 1992). Some attempts have already been made towards a joint model for death and malformation outcomes (Chen 1993). Kuk (2002) proposed a model for fetal response in developmental toxicity studies when the number of implants is dose-dependent.
We propose a Bayesian hierarchical modelling framework for the joint analysis of fetal death and malformation/weight among the viable fetuses. In a first step, we construct a model for the joint analysis of death and malformation. In a later step, we will extend this approach to include the weight of the viable fetuses.
Let N denote the total number of dams, and hence litters, in the study. For the ith dam ($i = 1, \ldots, N$), let m_i be the number of implants. Let r_i indicate the number of fetal deaths in cluster i. The number of viable fetuses, i.e., the litter size, is $n_i \equiv m_i - r_i$. The number of malformed fetuses of a dam is denoted z_i.
A joint model for the possible adverse fetal outcomes is developed using the underlying hierarchy of the data. In the first stage, a toxic insult may result in a fetal death. This effect of dose d_i on cluster i with m_i implants can be described using the distribution $f(r_i|m_i, d_i)$. We assume that

$$r_i \sim Binomial(p_{dth,i}, m_i)$$

with $p_{dth,i}$ the probability of a death fetus in litter i, depending on the dose. In the second stage, the fetuses that survived the entire gestation period are at risk of malformation. The effect of malformation of dose d_i on cluster i with n_i viable fetuses can be described using the distribution $f(z_i|n_i, d_i)$. We assume that

$$z_i \sim Binomial(p_{mal,i}, n_i)$$

with $p_{mal,i}$ the probability of a malformed fetus in litter i, depending on dose d_i. A joint model for the number of deaths and the number of malformations can be assessed by jointly modelling both stages.
To account for the litter effect, we assume a hierarchical model in which the probability of an adverse event in each litter come from a prior distribution. We assume the malformation and death probability p_i of any fetus in litter i to come from a beta distribution with mean π_i, i.e.,

$$p_{dth,i} \sim Beta(a_{1i}, b_{1i}) \qquad \pi_{dth,i} = \frac{a_{1i}}{a_{1i} + b_{1i}}$$
$$p_{mal,i} \sim Beta(a_{2i}, b_{2i}) \qquad \pi_{mal,i} = \frac{a_{2i}}{a_{2i} + b_{2i}}$$

Both the malformation and death probability are affected by dose, and can be modelled using appropriate link functions. We assume

$$\begin{aligned} \text{logit}(\pi_{dth,i}) &= \alpha_0 + \alpha_d d_i \\ \text{logit}(\pi_{mal,i}) &= \beta_0 + (\alpha_d + \beta_d) d_i, \end{aligned}$$

with a common parameter for the dose effect.
In a last step, we specify hyperprior distributions on the regression parameters $\alpha_0, \alpha_d, \beta_0$ and β_d. The hyperpriors chosen for this analysis were $N(0, 10^6)$. We expect these priors to have minimal influence on the final conclusions of our analysis.

4 Data Analysis

This article is motivated by the analysis of developmental toxicity of Ethylene Glycol (EG) in mice. EG is a high-volume industrial chemical with diverse applications. For instance, it can be used as an antifreeze, as a solvent in the paint and plastics industries, as a softener in cellophane, etc. The potential reproductive toxicity of EG has been evaluated in several laboratories. Price et al (1985) for example, describe a study in which timed-pregnant CD-1 mice were dosed by gavage with EG in distilled water. Dosing occurred during the period of organogenesis and structural development of the fetuses (gestational days 8 through 15). Table 1 shows the rate of malformed litters for each dose group and suggests clear dose-related trends for the rate of malformation. The mean litter size is also tabulated, and shows a decrease with dose.

TABLE 1. *Summary Data from an EG Experiment in Mice*

Dose (mg/kg/day)	Dams	Live	Litter Size (mean)	Malformations (%)
0	25	297	11.9	4.0
750	24	276	11.5	66.7
1500	22	229	10.4	81.8
3000	23	226	9.8	95.7

TABLE 2. *Risk Assessment for EG Study in Mice.*

Model	$q = 0.01$	$q = 0.05$
Joint	103	447
Malf	142	563
Death	340	1493

We define the combined risk due to a toxic effect as the probability that a fetus is death or a viable fetus is malformed. This risk can be expressed as

$$\begin{aligned} r(d) &= \text{P(death fetus} \mid d) + \text{P(viable fetus} \mid d) \times \text{P(malformed} \mid \text{viable}, d) \\ &= \pi_{dth} + (1 - \pi_{dth})\pi_{mal}. \end{aligned}$$

The benchmark dose is defined as the level of exposure corresponding to an acceptably small excess risk (q) over background, i.e., the dose satisfying

$$r^*(d) = \frac{r(d) - r(0)}{1 - r(0)} = q.$$

Table 2 shows the benchmark doses corresponding to the 1% and 5% excess risk. We also added the corresponding quantities, calculated from univariate risks (only malformation, or only death). The joint model yields more conservative doses.

References

Aerts, M., Geys, H., Molenberghs, G., and Ryan, L.M. (2002). *Topics in Modelling of Clustered Data*. Chapmann & Hall.

Ryan, L.M., Catalano, P.J., Kimmel, C.A., and Kimmel, G.L. (1991). Relationship between fetal weight and malformation in developmental toxicity studies. *Teratology*, **44**, 215–223.

Williams, P.L. and Ryan, L.M. (1996). Dose-Response Models for Developmental Toxicology. In: *Handbook of Developmental Toxicology of R.D. Hood (ed.)*, 636–666, New York: CRC Press.

Influence Assessments for Longitudinal Data in Linear Mixed Models

Yu Fei[1] and Jianxin Pan[2]

[1] Economics School, Yunnan University, Kunming, 650091, P.R.China
[2] Mathematics Department, Manchester University, Manchester, M13 9PL, U.K.
Email: jpan@maths.man.ac.uk

Abstract: The linear mixed model (LMM) has been widely used in longitudinal studies. Existing methods of influence assessments in LMM are mostly based on the likelihood function, in which the marginal covariance structure is usually assumed to be compound symmetry. These methods, however, may be very difficult to apply to the model with sophisticated but commonly encountered covariance structures such as AR(1) and ante-dependence. In this paper, we propose an alternative approach using Q-function, the conditional expectation of logarithm of the joint-likelihood in EM-algorithm. The effects of mis-specification of covariance structure on influence analysis are addressed and the relationship between subject- and observation-levels influences are considered.

Keywords: Influence; Linear mixed model; Longitudinal data; Q-function.

1 Introduction

The *linear mixed model* (LMM) is commonly used in longitudinal data analysis, which is defined by

$$Y = X\beta + Zu + \epsilon \tag{1}$$

where $Y = (Y_1', \cdots, Y_m')'$ is the $(n \times 1)$ $(n = \sum_{i=1}^{m} n_i)$ response vector of m subjects, $X = (X_1', \cdots, X_m')'$ is the $(n \times p)$ design matrix for fixed effects β, $Z = \mathrm{diag}(Z_1, \cdots, Z_m)$ is the $(n \times mq)$ design matrix for random effects $u = (u_1', \cdots, u_m')'$, $u \sim N(0, \mathcal{G})$, $\epsilon = (\epsilon_1', \cdots, \epsilon_m')'$ is the $(n \times 1)$ vector of random errors and $\epsilon \sim N(0, \mathcal{R})$. The random effects u_i are independent of the random errors ϵ_i, $\mathcal{G} = \mathrm{diag}(G, \cdots, G)$ where $G = G(\alpha)$, the $(q \times q)$ between-subject common covariance matrix, and $\mathcal{R} = \mathrm{diag}(R_1, \cdots, R_m)$ where $R_i = R_i(\gamma)$, the $(n_i \times n_i)$ within-subject covariance matrix, where α and γ are $r \times 1$ and $s \times 1$ parameters in G and R_i, respectively. The parameters of interest are $\theta = (\beta', \alpha', \gamma')'$.
When the ith subject is deleted, Eq.(1) reduces to $Y_{[i]} = X_{[i]}\beta + Z_{[i]}u_{[i]} + \epsilon_{[i]}$ where a vector/matrix with the index $[i]$ represents the associated vector/matrix with ith sub-vector/matrix removed. This model is called the subject-deletion model in longitudinal studies.

For the LMM with independent random effects and random errors, Benerjee and Frees (1997) developed an approach to quantify the overall impact of a subject on the modelling. Lesaffre and Verbeke (1998) addressed the issue of local influences, see also a recent review paper by Molenberghs and Verbeke (2001). When the covariance matrices $\mathcal{G}$ and $\mathcal{R}$ are of sophisticated structures such as AR(1) and ante-dependence, these methods may be too difficult to apply. In this paper, we propose an alternative approach based on Q-function, the conditional expectation of logarithm of the joint-likelihood in EM-algorithm, to identify influential subjects. The advantage of this technique is that it can be easily applied to sophisticated models. In Section 2, within the framework of LMM with compound symmetry covariance we compare the Q-based diagnostics to likelihood-based methods. In Section 3 we study the effects of mis-specification of covariance structures on influence assessments. Some further comments on influences in longitudinal studies are given.

2 Influence Assessments in LMM

2.1 Likelihood-based Approach

Let $l(\theta)$ and $l_{[i]}(\theta)$ be the log-likelihood functions under the full model (1) and the subject-deletion model, respectively. Denote $\hat{\theta}$ and $\hat{\theta}_{[i]}$ as the MLEs of θ associated with the two models. A vital issue in case-deletion influence assessments is to quantify the difference between $\hat{\theta}_{[i]}$ and $\hat{\theta}$.

Applying the Fisher-scoring algorithm to $l_{[i]}$, with $\hat{\theta}$ as the initial value of θ we obtain the one-step approximation estimate

$$\hat{\theta}_{[i]} = \hat{\theta} + \{-E\ddot{l}_{[i]}(\hat{\theta})\}^{-1}\dot{l}_{[i]}(\hat{\theta}), \tag{2}$$

where $\dot{l}_{[i]}(\hat{\theta})$ and $\ddot{l}_{[i]}(\hat{\theta})$ are the first- and second-derivatives of $l_{[i]}$, evaluated at $\hat{\theta}$. Although Eq.(2) provides an approximation of the difference $\hat{\theta}_{[i]} - \hat{\theta}$, the matrix $E\ddot{l}_{[i]}$ is subject-dependent and may cause intensive computations. For example, we have to compute m inverse matrices $\{-E\ddot{l}_{[i]}(\hat{\theta})\}^{-1}$ $(i = 1, 2, ..., m)$ when using Eq.(2). Instead, we propose to use $E\ddot{l}$ to replace $E\ddot{l}_{[i]}$, i.e.,

$$\hat{\theta}_{[i]} = \hat{\theta} + \{-E\ddot{l}(\hat{\theta})\}^{-1}\dot{l}_{[i]}(\hat{\theta}). \tag{3}$$

We can show that this approximation can be characterized by $O_p(n^{-2})$ under certain conditions.

Based on (3), a generalized Cook-type distance can be defined as

$$D_i = (\hat{\theta}_{[i]} - \hat{\theta})'\{-E\ddot{l}(\hat{\theta})\}(\hat{\theta}_{[i]} - \hat{\theta}) = [\dot{l}_{[i]}(\hat{\theta})]'\{-E\ddot{l}(\hat{\theta})\}^{-1}[\dot{l}_{[i]}(\hat{\theta})]. \tag{4}$$

If both the matrices G and R_i are of independent structures, i.e., $G = \sigma_u^2 I_q$ and $R_i = \sigma_\epsilon^2 I_{n_i}$, it can be shown that the matrix $E\ddot{l}(\hat{\theta})$ is of block-diagonal so that D_i can be easily calculated. When either G or R_i is of

other structures, however, no analytical form of D_i is available because the matrix $E\ddot{l}(\hat{\theta})$ may be too complicated.

2.2 Q-function-based Approach

In order to study influence analysis in LMM, we propose to use the Q-function in EM-algorithm to replace the likelihood-based methods. The Q-function for the model (1) is defined by $Q(\theta|\tilde{\theta}) = E\{\log f(Y,u)|Y,\tilde{\theta}\}$, the conditional expectation of the joint log-likelihood function of the responses Y and the random effects u, given the responses, where $\tilde{\theta}$ is the updated solution of θ in EM algorithm. Based on $\ddot{Q}$, the second-order derivative of Q with respect to θ, Zhu et al (2001) discussed influence assessments for incomplete data. However, the calculation of $\ddot{Q}$ for sophisticated models may be too difficult and the resulting generalized Cook's distance may have no clear interpretation. As an alternative we replace $\ddot{Q}$ with $E\ddot{Q}$, the expectation of $\ddot{Q}$, leading to the generalized Cook's distance:

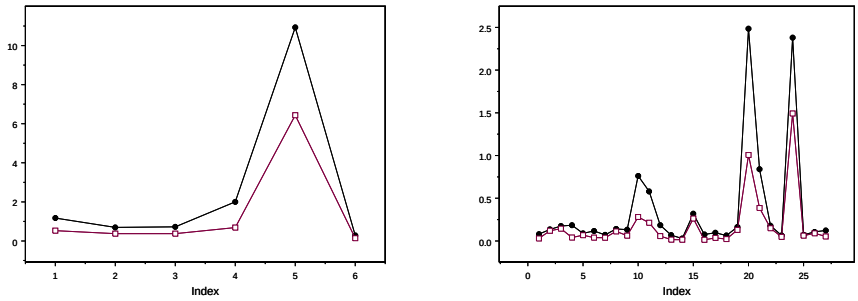
$$D_i^* = [\dot{Q}_{[i]}(\hat{\theta}|\hat{\theta})]'\{-E\ddot{Q}(\hat{\theta}|\hat{\theta})\}^{-1}[\dot{Q}_{[i]}(\hat{\theta}|\hat{\theta})], \tag{5}$$

where both the first-order derivative $\dot{Q}_{[i]}$ and the second-order derivative $\ddot{Q}$ are evaluated at the MLE $\hat{\theta}$. For the LMM (1), we find that the matrix $E\ddot{Q}$ is always block-diagonal whatever the covariance structure is. The generalized Cook's distance D_i^* in (5) hence can be decomposed into three components

$$D_i^* = D_{i\beta}^* + D_{i\alpha}^* + D_{i\gamma}^*, \tag{6}$$

where $D_{i\beta}^*$, $D_{i\alpha}^*$ and $D_{i\gamma}^*$ are the generalized Cook's distances corresponding to the fixed effects β, the between-subject covariance components α and the within-subject covariance components γ, respectively. In other words, the influence measurements for the three sets of parameters are mutually independent in this sense.

To measure how good the Q-based statistic D_i^* is, we compare it with the likelihood-based influence measurement D_i under the framework of LMM with $G = \sigma_u^2 I_q$ and $R_i = \sigma_\epsilon^2 I_{n_i}$ through analyzing two practical data sets. The first one is the Aerosol Data set (e.g., Beckman et al, 1987) and the second is the Dental Data set (e.g., Pan and Fang, 2002). For the Aerosol data Beckman et al (1987) identified that the 5th subject is the most influential subject. Pan and Fang (2002) analyzed the Dental data in terms of growth curve modelling and detected the 20th and 24th individuals as the two largest influential subjects. Figure 1(a) gives the index plots of D_i and D_i^* for the Aerosol data while Figure 1(b) displays the corresponding index plots for the Dental data, from which we see the performances of the two influence measurements are very close, implying that D_i^* is a good alternative to D_i.



(a) (b)

FIGURE 1. *The index plots of D_i and D_i^* for the Aerosol data (Panel (a)) and the Dental data (Panel (b)), where lines with dots and with empty boxes are Di and D_i^*, respectively.*

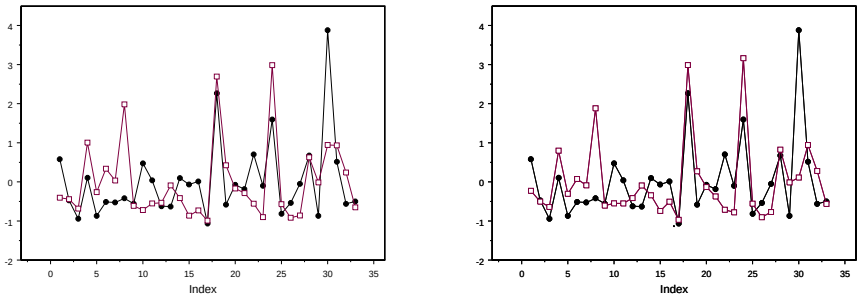
3 Effects of Covariance Structures

In practice, the study of the fixed effects may be the focus in longitudinal studies. A well-known result in the LMM is that a mis-specification of covariance structures may not affect the magnitude of estimates of the fixed effects (e.g., Pan and Fang, 2002). An interesting question is, does a mis-specification of covariance structures affect the fixed effects in terms of influence assessments?

To see this we analyzed Zerbe's Glucose data, in which the 30th subject was identified as the largest influential subject using growth curve modelling technique (Pan and Fang, 2002). BIC-based model selection criterion suggests that a linear mixed model with $G = \sigma_u^2 I_q$ and R_i =AR(1) is the best fitting, which is called Model 1. We then consider two models that have the same fixed effects and random effects to Model 1 but mis-specify the covariance structures for either G or R_i: (a)Model 2: $G = \sigma_u^2 I_q$ and $R_i = \sigma_\epsilon^2 I_{n_i}$ and (b) Model 3: G =AR(1) and $R_i = \sigma_\epsilon^2 I_{n_i}$. Although the mis-specification of covariance structures occurs, both models are not too far away from Model 1 in terms of BIC values (BIC=388.93, 418.21 and 419.64 for Models 1, 2 and 3, respectively).

For each model we compute the generalized Cook's distance $D_{i\beta}^*$ for the fixed effects β. Figure 2(a) gives the index plots of the statistic for Model 1 (line with dots) and Model 2 (line with empty boxes), while Figure 2(b) displays the comparison of Model 1 (line with dots) with Model 3 (line with empty boxes). Figure 2 shows that the 30th subject, the largest influential subject identified using Model 1, can not be detected as the largest influential subject using either Model 2 or Model 3. Instead, these two models identify the 24th subject as the most influential subject, which is in fact the third largest influential subject. We anticipate the reason that both

the mis-specified models identify the right influential subject, though not the largest influential one, is their closeness to the best model. In general, mis-identification of influential subjects may occur when mis-specifying covariance structures.



(a) (b)

FIGURE 2. *The index plots of $D^*_{i\beta}$ for Models 1 & 2 (Panel (a)), and for Models 1 & 3 (Panel (b))*

In summary, our proposed approach is a good alternative to likelihood-based influence methods. A mis-specification of covariance structures may lead to mis-identification of influential subjects. Correct specification of covariance structures is thus crucial for diagnostics purpose. We also studied the relationship between subject- and observation-levels influences. Further details will be reported in the oral presentation.

References

Banerjee, M. and Frees, E. W. (1997). Influence diagnostics for longitudinal models. *Journal of the American Statistical Association*, **92**, 999–1005.

Beckman, R. J., Nachtsheim, C. J., and Cook, R. D. (1987). Diagnostics for mixed-model analysis of variance. *Technometrics*, **29**, 413–426.

Lesaffre, E. and Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics*, **54**, 570–582.

Molenberghs, G. and Verbeke, G. (2001). A review on linear mixed models for longitudinal data, possibly subject to dropout. *Statistical Modelling*, **1**, 235–269.

Pan, J. X. and Fang, K. T. (2002). *Growth Curve Models and Statistical Diagnostics*. New York: Springer-Verlag.

Zhu, H., Lee, S. L., Wei, B. C., and Zhou, J. (2001). Case-deletion measures for models with incomplete data. *Biometrika*, **88**, 727–737.

Some Further Results on Time Series of Counts

Konstantinos Fokianos

[1] University of Cyprus, Department of Mathematics & Statistics, P.O. Box 20537, Nicosia 1678, Cyprus

Abstract: Regression models for time series of counts have been developed over the last years within the framework of generalized linear models methodology to take into account serial dependence that occurs so frequently in applications. Estimation, testing and prediction can be routinely carried out using standard conditional/partial likelihood methods under certain regularity conditions. The aim of this communication is to report some further results on this still evolving applied area by discussing an autoregressive moving average model for time series of counts. Several simulations enrich the theoretical results.

Keywords: Partial likelihood; Count data; Regression.

1 Introduction

Over the last years there has been a growing interest on regression models for time series of counts whose development has been facilitated by various applications arising frequently from different scientific disciplines such as finance, medicine, environmentrics to name a few, Kedem and Fokianos (2002). These regression models–often called "transitional", "conditional", and "Markov" models–are analyzed in almost all of the cases by partial likelihood or quasi-likelihood methods which allow for temporal or sequential conditional inference with respect to a filtration generated by all that is known to the observer at the time of observation. This enables very flexible conditional inference which takes into account autoregressive components, functions of past covariates, all forms of interactions among covariates, and more generally *time dependent random covariates*. Furthermore, the combination of partial likelihood and regression models for time series of counts provide a methodological sound framework where estimation, diagnostics, model assessment, and forecasting are implemented in a straightforward manner while the computation is carried out by a number of the existing software packages. These issues are addressed in the list of *desiderata* suggested in Davis et al (1999) and Zeger and Qaqish (1988) and have further examined in Kedem and Fokianos (2002).
The main objective of this work is to study an autoregressive moving average model for time series of counts whose moving average part depends

upon an unknown parameter, or possibly, parameters. Therefore estimation of the parameter/parameters involved in the moving average part of the model is necessary for regression fitting and prediction.
To fix notation, suppose that Y_t, $t = 1, 2, \ldots, N$ is a response time series of counts and let Z_{t-1}, $t = 1, \ldots, N$ be a p–dimensional vector of covariates which may includes past values of the process and/or any other auxiliary information. Under the above setup, statistical inference is mainly concerned with exploring the relationship between the expected value of the response and the covariates given the history. Thus, a methodologically sound analysis is based on the following regression model

$$\mu_t(\beta) = h(Z'_{t-1}\beta), \quad t = 1, \ldots, N, \tag{1}$$

where $\mu_t = \mathrm{E}[Y_t \mid \text{past}]$ and the *inverse link* $h(.)$ function maps a subset $H \subseteq R$ one–to–one onto $(0, \infty)$. The regression coefficients β are unknown and need to be estimated from the data. For instance, when h equals to the exponential function, then expression (1) leads to the so called log–linear model where

$$\log \mu_t(\beta) = Z'_{t-1}\beta. \tag{2}$$

In what follows, we are concerned with model (2). It is well known that inference about the regression parameters can be carried either by conditional or partial likelihood methods (Ch.6 of Fahrmeir and Tutz (1994) and Ch.4 of Kedem and Fokianos (2002)), or by the so called estimating equations approach, Zeger and Qaqish (1988).

2 A Moving Average Model

Consider model (2) and let $Z_{t-1} = (X_t, e_{t-1}, \ldots, e_{t-p})'$, where X_t denotes a multivariate auxiliary process and the random sequence $\{e_t\}$ is defined by

$$e_t = \frac{Y_t - \mu_t}{\mu_t^{\lambda}}, \quad t = 1, \ldots, N, \tag{3}$$

for $\lambda \geq 0$, see Davis et al (1999). Hence equation (2) becomes

$$\log \mu_t(\beta) = X'_t\gamma + \sum_{i=1}^{p} \theta_i e_{t-i},$$

where $\beta = (\gamma', \theta_1, \ldots, \theta_p)'$. A close examination of the above model resembles the well known ARMA models. In particular when $\lambda = 1/2$, then the random sequence $\{e_t\}$ reduces to the so called Pearson residuals and then Y_t possesses a number of nice properties. To mention only few

- model (2) can be used for prediction in a straightforward manner.

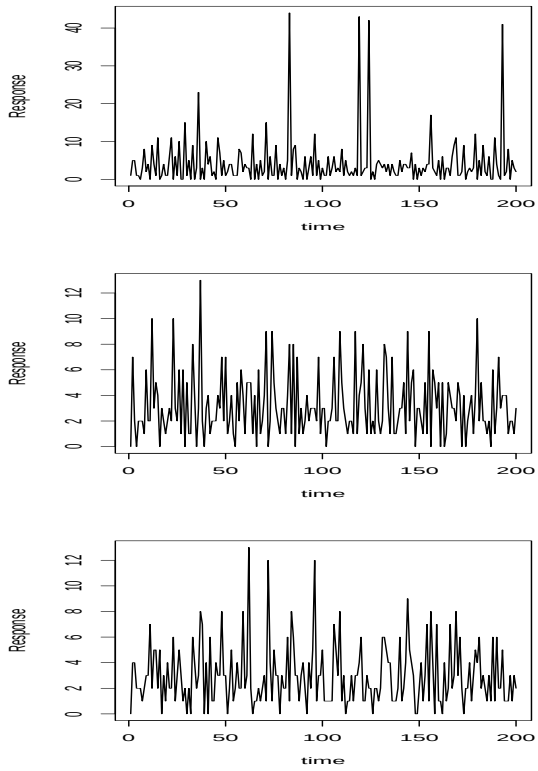


FIGURE 1. *Typical realizations of model (4) with* $\beta = (1, 0.5, -0.5)'$ *and* $N = 200$. *From top to the bottom:* $\lambda = 0, 0.5, 1$.

- model (2) takes into account serial dependence for estimating the parameter vector γ.

As an example consider the following model

$$\log \mu_t(\beta) = \gamma_1 + \gamma_2 \cos\left(\frac{2\pi t}{12}\right) + \theta_1 e_{t-1} \tag{4}$$

for $t = 1, \ldots, N$ and set $\beta = (\gamma_1, \gamma_2, \theta_1)'$ and $Z_{t-1} = (1, \cos(2\pi t/12), e_{t-1})'$. Then the log–linear model (2) is satisfied with this notation. Figure 1 display realizations of the observed time series of counts for different parameter values of both β and λ and points to a rather rapid oscillation for all λ–even though the value $\lambda = 0$ yields to a few extreme points.
Clearly, when the parameter λ is known, inference can proceed in a straightforward way according to the established theory since the vector $(e_{t-1}, \ldots, e_{t-p})'$ can be thought as additional covariates. However a problem arises when λ is not known and it is not well understood how estimation of f λ

affects the regression parameters. The aim of this contribution is to investigate

- Estimation of λ in (3).
- Prediction for count time series under joint estimation of β and λ.

References

Davis, R.A., Dunsmuir, W.T.M., and Wang, Y. (1999). Modelling time series of count data. In: *Asymptotics, Nonparametrics and Time Series*, 63–114. New York: Marcel-Dekker.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on generalized Linear Models*. Second Edition. Springer, New York: Springer-Verlag.

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

Zeger, S.L. and Qaqish, B. (1988). Markov regression models for time series: A quasi–likelihood approach. *Biometrics*, **44**, 1019–1031.

Sensitivity Analysis Based on Covariance Structures for Longitudinal Data with Dropout

M. Ganjali and M. Rezaee

[1] Dept. of Statistics, Faculty of Mathematical Sciences, Shahid Beheshti university, Theran, Iran.

Abstract: For the joint modelling of longitudinal continuous responses and dropout generalized Heckman model is used in order to see the influence of small perturbation of the elements of covariance structure on likelihood displacement. The perturbation from random dropout in the direction of informative dropout is considered for Mastitis data.

Keywords: Longitudinal data; Informative dropout; Sensitivity analysis; Generalized Heckman model.

1 Introduction

Recently joint modelling of response and non-response in cross-sectional and longitudinal data has been extensively used. Examples of such models are the selection model of Heckman (1979) and the dropout model of Diggle and Kenward (1994). In Diggle and Kenward's model dropout is at random (RD) if, given the previous outcome, it is independent of the current response and it is completely random dropout (CRD) if it neither depends on the previous nor on the current response. If dropout is not CRD or RD, it is informative dropout (ID). However, Diggle and Kenward's model rests on strong assumptions (discussion of Diggle and Kenward, 1994) and it has been so much suggested that an important way to use joint modelling is by means of sensitivity analysis (Verbeke and Molenberghs, 1997 and 2000 and the references mentioned there). Several tools have been discussed in the literature, such as the informal sensitivity analysis of Kenward (1998) and a local influence based approach as formal sensitivity analysis (Molenberghs et al, 2001) for assessment of the influence of a small modification of model components.

Molenberghs et al (2001) use Diggle and Kenward's model and the approach of Cook (1986) for measuring the influence of a small perturbation of the model components. In this paper we shall use the generalized selection model of Heckman (Crouchley and Ganjali, 2002, see also next Section) and the approach of Cook (1986) for measuring the influence of

a small perturbation of the model components for longitudinal data with dropout. The approach will be discussed for measuring the influence of a small perturbation of the covariance structure of the generalized Heckman model (GHM) on likelihood displacement (see Section 3). In Section 4, as an application, the Mastitis data will be used for assessing the influence of the perturbation from RD in the direction of NRD.

2 The Selection Model and its Generalization

Heckman (1979) proposed a joint model for a continuous response (y_i) and a sample selection mechanism. This model is defined by means of two equations,

$$\begin{aligned} R_i^* &= \alpha^T \mathbf{W}_i + v_i \\ y_i^* &= \beta^T \mathbf{X}_i + \varepsilon_i, \end{aligned}$$

where α and β are vectors of parameters, $\mathbf{W}_i$ and $\mathbf{X}_i$ are vectors of covariates, (v_i, ε_i) are i.i.d drawings from a bivariate normal distribution with zero means, variances $\sigma^2_{RR} = 1$, σ^2_{YY} and covariance σ_{RY}. It is assumed that y_i^* is observed only when $R_i^* > 0$. So let $y_i = y_i^*$ if $R_i^* > 0$ and $y_i = 0$ if $R_i^* \leq 0$, for $i = 1, ..., n$, where $y_i = 0$ is used to indicate a missing response. Also define $R_i = 1$ if $R_i^* > 0$ and $R_i = 0$ if $R_i^* \leq 0$, so that (y_i, R_i) constitute the observations for subject i.
The Heckman (1979) model is generalized to the situation of repeated responses with dropout by Crochley and Ganjali (2002). This is

$$\begin{aligned} R_{it}^* &= \alpha_t^T \mathbf{W}_{it} + v_{it} \\ y_{it}^* &= \beta_t^T \mathbf{X}_{it} + \varepsilon_{it}, \end{aligned}$$

where $t = 1, ..., T_i$, $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{iT_i})$ and $\mathbf{R}_i = (R_{i2}, ..., R_{iT_i})$. In this model, it is assumed that all the subjects at the start of the study are observed, i.e. $R_{i1} = 1, \forall\ i$. The observations for the subject i, take the form $(\mathbf{y}_i, \mathbf{R}_i) = \left(\left[y_{i1}^*, ..., y_{iT_{i-1}}^*, 0\right], [1, ..., 1, 0]\right)$, if dropout occurs and $(\mathbf{y}_i, \mathbf{R}_i) = \left(\left[y_{i1}^*, ..., y_{T_i}^*\right], [1, ..., 1]\right)$, if a subject is independently right censored by the observation plan at time T_i.
In this model $Var\,(\varepsilon_i) = \Sigma_{YY}$. Σ_{YY} is assumed to be unstructured so that $\mathrm{Var}(\varepsilon_{it}) = \sigma^2_{YY_t}$ and $cov\,(\varepsilon_{is}, \varepsilon_{it}) = \sigma_{YY_{s,t}}$. It is also assumed that the subjects are independent of each other, so that $cov\,(\varepsilon_{is}, \varepsilon_{i't}) = 0$ for $i \neq i'$ for all s and t. Also $Var\,(\mathbf{v}_i) = \Sigma_{RR}$, where $diag\,(\Sigma_{RR}) = 1$.

The Dropout Mechanism

When dropout occurs $y_{iT_i}^*$ is not observed. Little and Rubin (2002) note that for CRD the dropout process must be independent of both the observed responses $\mathbf{y}_{io}^* = \left(y_{i1}^*, ..., y_{iT_{i-1}}^*\right)$ and $y_{iT_i}^*$, while for RD the dropout process, conditional on $\mathbf{y}_{io}^*$, must be independent of $y_{iT_i}^*$.

If we let $f(.)$ denote a multivariate normal distribution then we have CRD if $f(\mathbf{R}_i^* \mid y_{T_i}^*, \mathbf{y}_{io}^*) = f(\mathbf{R}_i^*)$. We have RD if $f(\mathbf{R}_i^* \mid y_{T_i}^*, \mathbf{y}_{io}^*) = f(\mathbf{R}_i^* \mid \mathbf{y}_{io}^*)$. With either CRD or RD the joint probability of $(\mathbf{y}_i^*, \mathbf{R}_i^*)$ factors so that we can use $f(\mathbf{y}_{io})$ on its own for unbiased inference about β. If $f(y_{T_i}^* \mid \mathbf{R}_i^*, \mathbf{y}_{io}^*)$ does not simplify for CRD or RD we have informative dropout (ID).
Crouchley and Ganjali denotes the variance-covariance matrix Σ_{GH} for the elements $\left(\mathbf{y}_{io}^*, y_{iT_i}^*, \mathbf{R}_i^*\right)$ as

$$\Sigma_{GH} = \begin{bmatrix} \Sigma_{Y_oY_o} & \Sigma_{Y_oY_T} & \Sigma_{Y_oR} \\ \Sigma_{Y_TY_o} & \sigma^2_{Y_TY_T} & \Sigma_{Y_TR} \\ \Sigma_{RY_o} & \Sigma_{RY_T} & \Sigma_{RR} \end{bmatrix}.$$

and they found that if both $\Sigma_{Y_TR} = \mathbf{0}$ (missing at random) and $\Sigma_{Y_oR} = \mathbf{0}$ (observed at random) we have CRD, i.e. $\Sigma_{Y_TR|Y_o} = 0$. They also found that if

$$\Sigma_{Y_TR} - \Sigma_{Y_TY_o}\Sigma^{-1}_{Y_oY_o}\Sigma_{Y_oR} = \mathbf{0} \tag{1}$$

we have RD. So we can estimate a model under RD by imposing the constraint $\Sigma_{Y_TR} = \Sigma_{Y_TY_o}\Sigma^{-1}_{Y_oY_o}\Sigma_{Y_oR}$.
Consider as an example the case of a two period longitudinal data where the response at first time is observed for all individuals. In this case $\mathbf{y}_i = (y_{i1}, y_{i2})$, $\mathbf{R}_i = R_{i2}$ and let

$$\Sigma_{GH} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & 1 \end{bmatrix}$$

where $cov(y_{i1}, y_{i2}) = \sigma_{12}$ and $cov(y_{ij}, R_{i2}) = \sigma_{j3}$ for $j = 1, 2$. Consequently equation (1) for $\sigma_{22} > 0$, $\rho_{23} - \rho_{12}\rho_{13} = 0$ gives the conditions for ignorable dropout. This can occur when $\rho_{23} = \rho_{13} = 0$, which is completely random dropout (CRD) and when $\rho_{23} = \rho_{12}\rho_{13}$.

3 Likelihood Displacement (LD)

We are interested in the influence that selection exerts on the parameters of interest in GHM as our selection model. If $\Sigma_{Y_TR} - \Sigma_{Y_TY_o}\Sigma^{-1}_{Y_oY_o}\Sigma_{Y_oR} = 0$ we have RD process. In this case measurement model parameters can not be influenced by selection. Modification of $H = \Sigma_{Y_TR} - \Sigma_{Y_TY_o}\Sigma^{-1}_{Y_oY_o}\Sigma_{Y_oR}$ may lead to large difference in the model parameters. Let denote the log-likelihood function corresponding to the GHM by

$$\pounds(\gamma \mid H) = \sum_{i=1}^{n} \pounds_i(\gamma \mid H)$$

in which $\pounds_i(\gamma \mid H)$ is the contribution of the ith individual to the log-likelihood and $\gamma^T = (\beta^T, \alpha^T)$ is the parameter vector of measurement and

TABLE 1. *Results for mastitis data.*

	ID model		RD model		CRD model		IDWO	
Par.	Est	SE.	Est	SE.	Est	SE.	Est	SE.
β_0	5.765	0.0.09	5.765	0.090	5.765	0.009	5.598	0.086
η	0.315	0.138	0.719	0.107	0.719	0.107	0.617	0.434
ρ_{12}	0.470	0.087	0.581	0.071	0.581	0.071	0.727	0.054
ρ_{13}	-0.157	0.125	-0.149	0.013			-0.127	0.131
ρ_{23}	0.676	0.117					-0.127	0.934
σ_{11}	0.931	0.064	0.931	0.064	0.931	0.064	0.872	0.060
σ_{22}	1.274	0.113	1.138	0.088	1.138	0.087	1.044	0.100
α_0	0.634	0.130	0.667	0.131	0.667	0.132	0.645	0.133
-logL	308.771		311.389		312.013		275.998	

dropout mechanisms. Let shows $\pounds(\gamma) = \pounds(\gamma \mid H = \mathbf{0})$ where $\pounds(\gamma)$ is the log-likelihood function which corresponds to a RD model. Suppose H can be perturbed around $\mathbf{0}$. Let $\hat{\gamma}$ be MLEs for γ obtained by maximizing $\pounds(\gamma)$ and $\hat{\gamma}_H$be MLEs for γ obtained by maximizing $\pounds(\gamma \mid H)$. Now one can compare $\hat{\gamma}_H$ and $\hat{\gamma}$ as local influence. If $\hat{\gamma}_H$ and $\hat{\gamma}$ are similar, parameters estimates are robust to the perturbation of RD in the direction of ID. Strongly difference estimates shows that estimation procedure is highly sensitive to such modification. We can use the Cook's LD which defined as $LD(H) = 2[\pounds(\hat{\gamma}) - \pounds(\hat{\gamma}_H)]$. A graph of $LD(H)$ versus H can be used as the influence of perturbations. For two-period longitudinal data with dropout LD is $LD(H) = 2[\pounds(\hat{\gamma}) - \pounds(\hat{\gamma}_H)]$ where $H = \rho_{23} - \rho_{12}\rho_{13}$.

4 Mastitis Data: Model and Results

Mastitis can reduce the milk yield of infected animals. We shall use data of the total milk yield for 107 cows from a single herd, in two consecutive years, to investigate the relationship between yield and mastitis. Of 107 animals, 27 were infected in their second year which will be treated as missing. For these data the GHM is in the form

$$\begin{aligned} y_{i1}^* &= \beta_0 + \varepsilon_{i1}, \\ y_{i2}^* &= \beta_0 + \eta + \varepsilon_{i2}, \\ R_{i2}^* &= \alpha_0 + v_{i3}, \end{aligned} \tag{2}$$

where η gives the effect of time on the mean of the response. We delete the effect of explanatory variable, selected year, as the previous analysis

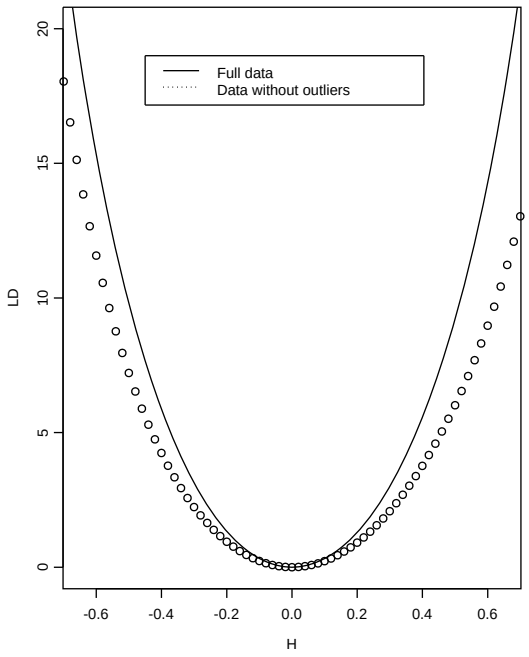


FIGURE 1. *Likelihood displacement against values of H.*

(Crouchley and Ganjali, 2002) shows no significant effect of this variable on responses. We used NAG (1996) routine E04UCF to obtain the likelihood displacements for these data.

Results from the GHM, System (2) for ID, RD, and CRD models are presented in Table 1.

We get an increase in deviance of 6.484 for 2 d.f. (p=0.039) for a test of CRD ($\rho_{13} = \rho_{23} = 0$) in System (2) and an increase in deviance of 5.236 for 1 df (p=0.022) for a test of RD ($\rho_{23} = \rho_{12}\rho_{13}$ in System (2)). Table 1 shows that, for the ID model, dropout is informative because of the stochastic dependency ($\rho_{23} = 0.676$) between the dropout process and the response in the second period. The value of ρ_{23} implies that a large value of the response in the second period (which may be missing) will increase the probability of being present in the second period. All the models give a significant change in mean response in the second period, but the CRD

and RD model overestimates it.
Using Pearson residuals Crouchley and Ganjali (2002) find 3 outliers in responses (cows 4, 5, 66). Deleting these observations show no sign of ID (see results of IDWO in Table 1). Figure 1 shows the LD against different values of H for full data and data without outliers.
In Figure 1, as it can be seen, there is no strong difference between LD for full data and LD for data without outliers. This suggests that only some outliers are the cause of ID in these data.

Acknowledgments: Many Thanks to Shahid Beheshti university for the financial support.

References

Cook, R.D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society, Series B*, **48**, 133–169.

Crouchley R. and Ganjali M. (2002). The common structure of several models for non-ignorable dropout. *Statistical Modelling*, **2**, 39–62.

Diggle, P. J. and Kenward, M. G. (1994). Informative Drop-out in longitudinal data analysis. *Applied Statistics,* **43**, 49–93.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, **47**, 153–161.

Kenward, M. G. (1998). Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. *Statistics in Medicine*, **17**, 2723–2732.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. New York: John Wiley

Molenberghs, G., Verbeke, G., Thijs, H., Lesaffre, E., and Kenward, M.G. (2001). Influence analysis to assess sensitivity of the dropout process. *Computational Statistics and Data Analysis*, **37**, 93-113.

Verbeke, G. and Molenberghs (1997). *Linear Mixed Models in Practice: A SAS-Oriented Approach.* Lecture Notes in Statistics 126, New York: Springer-Verlag.

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

Bivariate Marker Development with Censored Values and Informative Dropout

Ronald B. Geskus[1]

[1] Municipal Health Service, Cluster of Infectious Diseases, Nieuwe Achtergracht 100, 1018 WT Amsterdam, The Netherlands and Dept. of Medical Statistics, Molecular Epidemiology and Clinical Informatics, Leiden University Medical Center, P.O.Box 9604, 2300 RC Leiden

Abstract: Performance of a selection model and a random effects model are compared under the presence of informative censoring, using real as well as simulated data. The bias in the parameter estimates from a random effects model when informative dropout is present is small in our data set. This result is confirmed by the simulation study. Larger differences are seen in the estimates of the individual random effects.

Keywords: Informative dropout; Repeated measurements; Survival; HIV/AIDS markers.

1 Introduction

Markers are internal host factors that represent the current disease or recovery status of an individual. Identification of markers and modeling their development provides information on the disease mechanism. After infection with the human immunodeficiency virus (HIV), the number of CD4 lymphocytes decreases, finally leading to severe immunodeficiency and AIDS. Recently, assays have become available to quantify the HIV-RNA level in HIV infected individuals. Marker data that are collected during follow-up are usually missing to some extent since the value of the marker is associated with the risk of dropout (death).
Numerous papers have been published that model the development of CD4 count after HIV seroconversion. Some papers included data on AIDS diagnosis in order to correct for informative dropout via selection models (e.g. Berzuini and Larizza (1996), Wulfsohn and Tsiatis (1997)) or mixture models (Pawitan and Self (1993)). Others investigated the amount of bias caused by informative dropout when dropout is ignored and only a random effects model is used (Faucett and Thomas (1996), Touloumi et al (1999)).
In the medical literature, the so-called set point theory has been the favourite model to describe the development of HIV-RNA level: after a short peak in the first few months after infection, levels tend to be stable until another rise occurs close to AIDS diagnosis. The development

of HIV RNA has been modeled in a couple of papers, taking account of the fact that the assay used has a lower detection limit. Some only model the repeated measures (Hughes (1999)), others also include the correction for informative dropout via selection models (Jacqmin-Gadda et al (2000), Lyles et al (2000)). The bivariate development of both markers, without considering informative dropout, has also been modeled (Boscardin et al (1998)).
Many studies have investigated the effects of age and some genetic mutations on AIDS progression via a Cox model. By combining a model for the effects of the cofactors on the marker development and the effects of the markers on AIDS risk via a selection model, insight is gained into the causal mechanisms of these cofactors (Taylor et al (2000)).
We use a selection model to investigate the ways in which age and the genetic cofactors influence progression to AIDS. Emphasis will be on the bias in the parameters of the longitudinal part that may be introduced when a random effects model is used that neglects the informative dropout. The results are compared with the results from a small simulation study and with the results from an ordinary least squares model that considers each value as independent.

2 Materials and Methods

We used data from two different cohort studies, the Amsterdam Cohort Study among homosexual men (N=126) and the French SEROCO Cohort Study (N=274). The Amsterdam Cohort Study among homosexual men was started in 1984. We only used information until the date that administration of highly active anti-retroviral therapy (HAART) became widespread in the Netherlands (July 1st, 1996). Follow-up data from hospitals was included as well. All laboratory measurements were done in one laboratory. The French SEROCO cohort was started in 1988. HIV-infected adults from 17 hospitals and a network of private practitioners have been enrolled. In the analysis, only homosexual men were included. Information until February 1st, 1996 was used, when HAART became widely available in France. Marker measurements originate from 19 different laboratories.
We used a selection model, in which the marker development is modeled and the AIDS risk is modeled conditional on the fitted marker values. The model for the marker development is

$$\begin{pmatrix} \mathrm{CD4}(t_{ij})^{1/3} \\ \log \mathrm{RNA}(t_{ij}) \end{pmatrix} = \begin{pmatrix} a_1^i + b_1^i\, t_{ij} + c_1^i\, t_{ij}^2 + \theta_1^{\mathrm{cal}} + \varepsilon_1(t_{ij}) \\ a_2^i + b_2^i\, t_{ij} + \gamma_2\, t_{ij} I_{(t_{ij}<0.5)} + \theta_2^{\mathrm{cal}} + \varepsilon_2(t_{ij}) \end{pmatrix},$$

with

$$\begin{aligned} (a_1^i, b_1^i, c_1^i, a_2^i, b_2^i)^T &\sim \mathcal{N}((\alpha_1(l), \beta_1, \gamma_1, \alpha_2^{\mathrm{site}}, \beta_2)^T, \Sigma), \\ \alpha_1(l_F) &\sim \mathcal{N}(\mu_F, \sigma^2_{\mathrm{lab}}), \end{aligned}$$

$$\begin{aligned}\alpha_1(l_A) &= \mu_A,\\ \varepsilon_1(t_{ij}) &\sim \mathcal{N}(0,\sigma^2_{\text{cal}}),\\ \varepsilon_2(t_{ij}) &\sim \mathcal{N}(0,\tau^2),\end{aligned}$$

and $\varepsilon_k(\cdot)$ independent. For the relation between the markers and the AIDS risk, we used a time-dependent Cox model

$$\lambda(t) = \lambda_0(t)\exp\{\text{fitted(CD4)} + \text{fitted(RNA)} + \text{cofactors}\}.$$

The fitted marker values are obtained by combining the fixed and random effects. The effect of the fitted cube root of CD4 count on AIDS risk was modelled via a linear spline with knots at the values five and seven. The fitted log RNA level was fitted linearly.
Since laboratory methods improved over time, we allowed for a calendar time effect for error variance and CD4 level. For Amsterdam, changes occurred in 1988 and 1992. For France, where this was dependent on the laboratory, we allowed for a change in 1990 and included a random laboratory effect. Since HIV-RNA level was determined retrospectively, we modeled an effect of calendar time on the level (representing the effect of storage of frozen samples and the increasing availability of treatment), but no effect on the variance of the measurement error.
We used a Bayesian approach to parameter estimation, starting with non-informative priors for the parameters. Posterior distributions were obtained via Markov Chain Monte Carlo techniques, using the WinBUGS package. Three chains with different sets of initial values were generated. In order to reduce posterior correlations, we used hierarchical centring in the parametrisation (Gelfand et al (1995)).
Results from the joint model were compared with the results from a random effects model that only incorporates the marker trajectories without correction for dropout due to AIDS.

3 Results

In total, we had 6761 CD4 records. The number of measurements per person ranged from 1 to 59 (median 14). For HIV-RNA level, we had 3807 records, ranging from 1 to 55 per person (median 8). Of the HIV-RNA records, 344 were below the detection limit.
Hierarchical centring greatly improved convergence. Only about 4000 iterations were needed instead of 50000 if no centring was done. Moreover, updating was done about three times faster.
The selection model and the random effects model give more or less similar parameter estimates for the population effects. Difference in parameter estimates, relative to the width of the 95% credibility interval of the parameter under the selection model, remains below 12%. The largest differences are

TABLE 1. *Parameter estimates and coverage probabilities of the 95% credibility intervals (in brackets). (ID=Informative dropout, MAR=Missing At Random)*

	random effects model		ordinary least squares model	
	$\overline{\alpha}$	$\overline{\beta}$	$\overline{\alpha}$	$\overline{\beta}$
ID	9.59 (0.94)	-0.57 (0.92)	8.62 (0.00)	-0.21 (0.000)
MAR	9.60 (0.95)	-0.58 (0.95)	9.51 (0.23)	-0.59 (0.086)

seen for some of the parameters from the covariance matrix of the random effects (σ_{b_2} : -11%, σ_{c_1} : -6.1%, ρ_{b_2,c_1} : 7.9%) and the parameters for the RNA development (calendar period effect France: -8.5%; calendar period effect Amsterdam 1988-1992 relative to before 1988: 12%; effect age on RNA slope: -7.7% and RNA slope parameter β_2: 5.1%). Relative biases in all other effects remain below 5%. Larger differences are seen in the individual random effects for individuals who had few records and a long period between the last record and the moment of censoring or AIDS.
After the initial drop, HIV-RNA load increases again at the population level. The same pattern is seen for the selection model and the random effects model. However, the ordinary least squares model gives highly biased results: after the initial drop, HIV-RNA remains at a stable level.

4 A Simulation Study

We did a simulation study in order to investigate the amount of bias in the random effects model and in the ordinary least squares model when the probability of dropout depends on a person's marker trajectory. We restricted to the development of one marker (CD4 count), without covariates. Individual intercepts a_i and slopes b_i follow a bivariate normal distribution, with mean $(\alpha, \beta) = (9.6, -0.58)$ and $\sigma_a = 1.48$, $\sigma_b = 0.45$, $\rho_{a,b} = -0.495$. For each person, a random censoring time is generated from a uniform distribution on $[0, 20]$. The hazard of the event of interest is given by $\lambda(s) = \lambda \exp\{\gamma \times (a^i + b^i s)\}$, with $\lambda = 1.8$ and $\gamma = -0.5$. The time span between subsequent observations is drawn from an exponential distribution with mean 0.25 years. Two thousand samples were generated, each sample containing data from 400 persons. We also generated one thousand samples in the situation that dropout only occurred through the censoring mechanism. Results are summarized in Table 1. Again we see that the random effects model performs well under our informative dropout mechanism.

5 Conclusions

The bias in the parameter estimates from a random effects model when informative dropout is present is small in our data set used. This result is

confirmed by the simulation study. It contradicts earlier results (Faucett and Thomas (1996), Touloumi et al (1999)), but may be explained by the large number of records per individual and the good follow-up until AIDS or censoring. The largest bias, although still small, is found in the parameters that describe HIV-1 RNA development and in the covariance matrix of the random effects. Also, larger biases are found in the individual random effects. The set point theory of viral load development seems to be an artefact caused by the frequent use of an ordinary least squares model to describe HIV-RNA development.

References

Berzuini, C. and Larizza, C. (1996). A unified approach for modeling longitudinal and failure time data, with application in medical monitoring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **18**, 109–123.

Boscardin, W.J., Taylor, J.M.G., and Law, N. (1998). Longitudinal models for AIDS marker data. *Statistical Methods in Medical Research*, **7**, 13–27.

Faucett, C.L. and Thomas, D.C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine*, **15**, 1663–1685.

Gelfand, A.E., Sahu, S.K., and Carlin, B.P. (1995). Efficient parametrisations for normal linear mixed models. *Biometrika*, **82**, 479–488.

Hughes, J.P. (1999). Mixed effects models with censored data with application to HIV RNA levels. *Biometrics*, **55**, 625–629.

Jacqmin-Gadda, H., Thiébaut, R., Chêne, G., and Commenges D. (2000). Analysis of left-censored longitudinal data with application to viral load in HIV infection. *Biostatistics*, **1**, 355–368.

Lyles, R.H., Lyles, C.M., and Taylor, D.J. (2000). Random regression models for human immunodeficiency virus ribonucleic acid data subject to left censoring and informative drop-outs. *Applied Statistics*, **49**, 485–497.

Pawitan, Y. and Self, S. (1993). Modeling disease marker processes in AIDS. *Journal of the American Statistical Association*, **88**, 719–726.

Taylor, J.M.G., Wang, Y., Ahdieh, L., Chmiel, J.S., Detels, R., Giorgi, J.V., Kaslow, R., Kingsley, L., and Margolick, J. (2000). Causal pathways for CCR5 genotype and HIV progression. *Journal of Acquired Immune Deficiency Syndromes*, **23**, 160–171.

Touloumi, G., Pocock, S.J., Babiker, A.G., and Darbyshire, J.H. (1999). Estimation and comparison of rates of change in longitudinal studies with informative drop-outs. *Statistics in Medicine*, **18**, 1215–1233.

Wulfsohn, M.S. and Tsiatis, A.A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330–339.

Multivariate Modeling of Computer Cache Rates via Regression Model Integration

Ilya Gluhovsky[1]

[1] Sun Microsystems Laboratories, 2600 Casey Ave MTV29-120, Mountain View, CA 94043, USA

Abstract: This work is motivated by the need of knowing computer cache rates over a wide range of architectural configurations while being able to simulate only a small fraction of them. We present methodology for fitting multivariate models to available cache rate data. The models must be smooth and possess two features. First, they interpolate (replicate) the data set itself. Second, they produce meaningful extrapolation beyond the range of the simulation data along several dimensions simultaneously corresponding to large cache configurations. This is achieved by gradually transitioning from an interpolation model to a smooth extrapolation model via a set of intermediate models of varying smoothness.

Keywords: Nonparametric regression; Smoothing; Extrapolation; Cache simulation.

1 Introduction

We consider the following regression problem. Given sample pairs $(\underline{x}_i, y_i)$ that leave large portions of the domain region unexplored, we would like to estimate the functional relationship between x and y fulfilling the following three requirements. Let f denote the estimate, $\mathcal{X}$ be the domain region of interest and $\mathcal{C} \subset \mathcal{X}$ be the data region. In the absence of "holes" in the data, it can be defined as the convex hull of the $\underline{x}_i$.
Requirement 1. $f(\underline{x}_i) = y_i$. This is the interpolation property.
Requirement 2. Over the extrapolation region $\mathcal{X} \setminus \mathcal{C}$, f is the smoothest possible function that still captures the trends in the data.
Requirement 3. f is smooth over $\mathcal{X}$.
Req. 2 implies that f should be very smooth over the extrapolation region while *Req. 1* stipulates that f must be rough enough to pass through all the data points. Therefore, f has to make a smooth transition achieved by gradually increasing the smoothness of f as we move away from the data. If the data are sampled evenly within $\mathcal{C}$ (no holes), the smoothness will be increased as we move away from $\mathcal{C}$. Constructing models of variable smoothness has been widely considered (Fan and Gijbels (1995), Ruppert (1997), Schimek (2000)). Frequently, one selects the amount of smoothness locally based on a criterion, such as the mean squared error (MSE). The primary

distinguishing feature of modeling cache rate data is the need for extrapolation. Smoothness selection is hard for such models because one is unable to estimate the model bias over $\mathcal{X} \backslash \mathcal{C}$ (nor variance for heteroscedastic models) deeming the local MSE approach infeasible. A secondary feature is that the data set is noiseless, thus, one need not consider smoother models at the expense of breaking *Req. 1.* It will be seen, however, that a similar method would work for noisy data as well for building a conditional expectation estimate f over $\mathcal{X} \setminus \mathcal{C}$ after standard smoothing is carried out over $\mathcal{C}$.
Our motivation for considering this problem is computer system performance estimation. Computer system models take as inputs cache rate data and output performance estimates as well as resource utilizations that allow one to identify bottlenecks. A cache rate is the frequency of a particular memory event. A (single level cache) system configuration is parameterized by the number of processors, cache size, and some other cache attributes. (Hennessy (2003)). A typical dataset consists of four to six predictors (the configuration parameters) and a continuous response (the cache rate). Caches can usually be simulated over a regular design covering the set of *simulatable* architectures. This would be the set $\mathcal{C}$. Gluhovsky (2003) discusses cache simulation constraints. However, performance analysis is usually required over a considerably broader domain requiring extrapolation. For example, we were only able to simulate configurations with up to 16 processors, while we are interested in designing machines with as many as 256 (logical) processors!

2 Building the Model

The first step in building model f is fitting a set of models $f_0, \cdots, f_M$ of different constant smoothness. The models are indexed with increasing smoothness. The idea then is to use the roughest interpolation model f_0 over most of $\mathcal{C}$ and switch to progressively smoother model as we move away from $\mathcal{C}$.
The emphasis of this work is on integrating the models of different smoothness. Our hope is that the ideas are applicable to a variety of modeling methodologies. To be concrete, for f_0 we chose an interpolating thin-plate spline. It is the smoothest function that passes through all the data points, thus, satisfies *Req 1.* Details of the smoothness criterion and fitting thin-plate splines can be found in Green and Silverman (1994). It is well known that the behavior of thin-plate splines is not suitable for extrapolation. Therefore, we move to a smoother set of models. In this work, we chose $f_1, \cdots, f_M$ to be additive models with bivariate interactions. They are fitted via backfitting (Hastie and Tibshirani (1990)) using a locally weighted running-plane smoother (Hastie and Tibshirani (1990), Cleveland and Devlin (1988)) with the same weighting kernel bandwidth for each of the additive components. Those bandwidths increase from f_1 to f_M (thus f_M is the

smoothest model) and are chosen to be equally spaced on the logarithmic scale. Our goal now is to keep f_0 over and around the data points, use f_M for much of the extrapolation region and use the intermediate models to facilitate the transition.

First, let us quantify *Req. 2*. We have to limit the smoothness of f from above; a model that is too smooth at the expense of losing the trend of the data would not define useful extrapolation. For a straightforward measure of the goodness of fit of f', we chose a high, say, $\alpha = .9$-quantile of the absolute residuals $|f'(\underline{x}_i) - y_i|$, where f' has constant smoothness. Denote the α-quantile by r_α. Then we constrain all models f_m to satisfy

$$r_\alpha \{|f_m(\underline{x}_i) - y_i|\} \leq \gamma \tag{1}$$

for some threshold γ.

The idea behind model integration is the following. As we move away from the data, rougher models become unstable and we observe wider discrepancy between the $M+1$ fits. Observing the discrepancy at $\underline{x}$ that is higher than that over the data region points towards using a smoother model. We would like to catch the transition early enough, so that jumping to a smoother model is smooth. On the other hand, we do not want to sacrifice the quality of the fit near the data by using too smooth a model.

Define the model m instability of the fit measure $s(\underline{x}, m)$ at $\underline{x} \in \mathcal{X}$ via

$$s(\underline{x}, m) = \text{range}\{f_k(\underline{x}) \ : \ \ k \geq m\}.$$

If $s(\underline{x}, m)$ is small and models $f_k, k \geq m$ produce similar fits at $\underline{x}$, this is an indication that the variability of the estimates due to extrapolation is not much of as issue for model m. It is possible that we are not flexible enough to pick up the trend of the data as well as we could. On the other hand, a large $s(\underline{x}, m)$ indicates that any potential gain in picking up the trend of the data is likely to be offset by loose control of the behavior of f_m.

Define $R(m) = r_\alpha \{s(\underline{x}_i, m)\}$, the α-quantile of the instability measures over the data points for, say, $\alpha = .9$. Let

$$p(\underline{x}, m) = s(\underline{x}, m) + (R(0) - R(m)) \tag{2}$$

define the penalty for using model m at $\underline{x}$. Its interpretation is that we penalize for the instability of the fit by $s(\underline{x}, m)$, but we discount it by $R(m)$ as the part being consistent with an interpolation problem. Adding $R(0)$ makes both summands nonnegative, as $R(m)$ decreases with m. In what follows let $R(m)$ be redefined as $R(m) \leftarrow R(0) - R(m)$.

We also considered another measure for whether or not a location $\underline{x}$ should be treated as part of extrapolation based on how far from the data points $\underline{x}$ is. Rather than examining the geometry of the data set, which is generally not tractable in several dimensions, the idea is to fit a smooth model to the indicator vector $1_{\{\underline{x}_i\}}$ in $\mathcal{X}$ of the data points (it is zero over $\mathcal{X} \setminus \{\underline{x}_i\}$).

A smooth function I that approximates this indicator is obtained by fitting a logistic locally weighted running plane model (Hastie and Tibshirani (1990)). To use $I(\underline{x})$ in the penalty definition (2), we opted for

$$p(\underline{x}, m) = s(\underline{x}, m) + r(m) + (1 - (c_0 - I(\underline{x}))_+/c_0)(R(m) - r(m)) \qquad (3)$$

for a given c_0, where $r(m)$ is defined similarly to $R(m)$ except for a smaller α. c_0 is taken to be, say, $r_{.1}\{I(\underline{x}_i)\}$. Thus, we lower the tolerance to instability of the fit if we are told by I to expect to use a smoother model. The reason to involve c_0 is to keep the original tolerances $R(m)$ over most of the data points.

Due to space limitations, we can only give a flavor of the last piece of the procedure. Observe that while the penalties (3) of two different models at some $\underline{x}$ may be similar, their fits may be rather different. If two different models are used at nearby locations, it may lead to a roughness in f. If the two models have similar penalties, we would prefer to discourage such behavior by incurring a slightly larger combined penalty, but using the same model at both locations. In general, we expect that as we move away from the data, the models used gradually become smoother, thus, defining contiguous regions where the same model is used. The details will be presented elsewhere.

3 Cache Rate Data Model

Space limitations allow us to only show a glimpse at the procedure. Figure 1 presents a slice through a 4-dimensional surface f for the total cache miss rate varying the number of processors with the other three arguments fixed at some values. There are four data points within this slice corresponding to processor levels 0, 2, 3, and 4 shown as small dots. The small dots (including those just mentioned) depict the fit given by f_0. The lines are smooth additive fits $f_1, \cdots, f_6$. The circles are the final model f. As can be seen, the procedure chose the second roughest model f_1 at level 1 that is close to the data and the smoothest model f_6 at levels 7, 8, and 9 far from the data. The transition was facilitated by two intermediate models. We can observe a widening discrepancy between the fits as we move away from the data and a particularly loose behavior of f_0 in the extrapolation region.

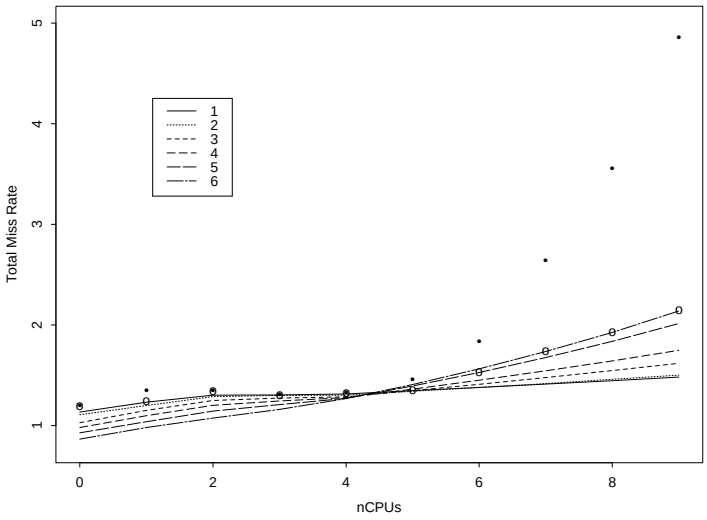


FIGURE 1. *Slice through f showing all the models.*

References

Cleveland, W.S. and Devlin, S.J. (1988). Locally-weighted regression: an approach to regression analysis by local fitting. *Journal of American Statistical Association*, **83**, 597–610.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth an spatial adaptation. *Journal of the Royal Statistical Society, Series B*, **57**, 371–394.

Gluhovsky, I. and O'Krafka, B.W. (2003). Comprehensive multiprocessor cache rate generation using multivariate models. Submitted.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall.

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

Hennessy, J.L. and Patterson, D.A. (2003). *Computer Architecture: A Quantitative Approach. Third Edition.* Morgan Kaufmann Publishers.

Kirkpatrick, S., Gelatt, C.D., and Vecchi, M.P. (1983). Optimization by simulated annealing. *Science*, **220**, 671–680.

Ruppert, D. (1997). Empirical bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of American Statistical Association*, **92**, 1049–1062.

Schimek, M.G. (ed.) (2000). *Smoothing and Regression*. New York: Wiley.

A Latent Variable Model for Joint Analysis of Repeatedly Measured Ordinal and Continuous Outcomes

R. Gueorguieva[1] and G. Sanacora [2]

[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
[2] Department of Psychiatry, Yale University School of Medicine, 34 Park St. New Haven, CT 06519, USA

Abstract: Biomedical researchers are often interested in estimation and testing of treatment effects on multiple outcome variables. A motivating example for this paper is a randomized clinical trial of two depression treatments in which an ordinal and a continuous measure of depression severity are collected repeatedly over 7 weeks. We formulate a latent variable mixed model for joint analysis of the two outcomes and compare results between joint and separate analysis of the response varaibles. Maximum likelihood estimation via adaptive Gaussian quadrature is used for estimation and inference. Likelihood-based methods are used for model comparison. Bias and efficiency comparisons between joint and separate fitting of the response variables are performed via simulations.

Keywords: Latent variable; Gaussian quadrature; Multivariate response; Repeated measures.

1 Introduction

In randomized clinical trials and observational studies medical researchers often measure multiple outcome variables and analyze these responses separately. Traditionally in such situations either no correction or the very conservative Bonferroni correction for multiple tests is applied. Since the outcome variables often correspond to the same or related latent processes, models with correlated or shared random effects can be fitted to several related outcome variables. This approach can keep the alpha level closer to the nominal level, may improve efficiency of treatment effect estimates, and may provide additional information about the relationship between variables. A number of authors have considered such models for cross-sectional data or for repeatedly measured data on responses of the same type. The case of multiple outcomes and repeated measures is complicated since two types of correlations must be taken into account: correlations between mea-

TABLE 1. *Average HAMD and CGI scores for the two treatment groups over time in the randomized clinical trial of depression.*

	HAMD		CGI	
Week	Active augmentation	Active	Active augmentation	Active
0	29.27	31.08	4.12	4.30
1	23.27	26.13	3.60	3.88
2	18.43	20.91	3.14	3.52
3	15.57	19.82	2.90	3.50
4	14.35	18.61	2.65	3.39
5	10.58	14.25	2.58	2.89
6	8.89	11.68	2.00	2.47

surements on different variables and correlations between measurements on the same variable within cluster or subject.

In this paper we formulate a latent-variable mixed model for a combination of ordinal and continuous outcomes measuring the same underlying process over time. We use maximum-likelihood estimation for model fitting and likelihood methods for model comparison. The motivating dataset is from a double-blind randomized clinical trial of two depression treatments (Sanacora et al., 2003). Fifty subjects with a diagnosis of major depression are randomly assigned to receive either active or active augmentation treatment. Hamilton depression scale ratings (HAMD) and clinical global impression (CGI) ratings are obtained weekly over a period of 6 weeks. Both variables measure depression severity. The hypothesis of interest is whether the experimental treatment group demonstrates faster improvement than the standard treatment group. The HAMD score is best treated as continuous measure and can be assumed to be normally distributed based on normal probability plots. The CGI is an ordinal variable measuring severity of illness on a scale from 1 to 7, with 1 indicating "normal, not at all ill" and 7 indicating "among the most extremely ill patients".Since there is only one value in categories 6 and 7, categories 5, 6 and 7 are combined in one. Means for each treatment group over time are shown in Table 1.

2 Model Definition and Properties

Let y_{ij1} and y_{ij2} denote the continuous and the ordinal outcome respectively, measured on the i^{th} subject $i = 1, ...I$ at the j^{th} time point, $j = 1, ...J$. Let also l_{ij} be the true unobserved depression status for the i^{th} subject at time j.

The model is defined by the following equations:

$$
\begin{aligned}
y_{ij1} &= \beta_0 + \beta_1 l_{ij} + \epsilon_{ij1}, \\
y_{ij2} &= \text{c} \quad (\text{if } \tau_{c-1} < l_{ij} + \epsilon_{ij2} < \tau_c,\ \text{c=1,...5},\ \tau_0 = -\infty,\ \tau_5 = +\infty), \\
l_{ij} &= \mathbf{x}_{ij}^T \boldsymbol{\gamma} + \mathbf{z}_{ij}^T \mathbf{b}_i + \epsilon_{ij}.
\end{aligned}
$$

Here β_0 is a shift parameter, β_1 is a factor loading parameter, the θ's are unknown thresholds. As usual $\mathbf{x}_{ij}^T$ and $\mathbf{z}_{ij}^T$ are covariate vectors, $\boldsymbol{\gamma}$ is a parameter vector describing the effect of covariates on the latent response and $\mathbf{b}_i$ are subject-specific random effects. The fixed effects in the latent variable model are treatment, linear time, quadratic time, treatment by linear time, treatment by quadratic time. We assume a random intercept and a random slope. Let also $\epsilon_{ij} \sim i.i.d.N(0, \sigma^2)$, independent of $\epsilon_{ij1} \sim i.i.d.N(0, \sigma_1^2)$, $\epsilon_{ij2} \sim i.i.d.N(0, \sigma_2^2)$ and of the random effect vector $\mathbf{b}_i \sim i.i.d.N(0, \Sigma)$. For identifiability we set $\gamma_0 = 0$ and $\sigma_2 = 1$.
If the underlying continuous outcome is observed this model reduces to the model of Roy and Lin (2002) with no variable-specific random intercepts. If each outcome is considered separately, the above formulation corresponds to a linear mixed model for the normal response and to a correlated probit model for the ordinal response. The proposed model can also be rewritten as a generalized latent and linear mixed model (GLAMM, Rabe-Hesketh, Pickles and Skrondal (2001)) and fitted using the gllamm function in Stata. Our model provides a test for a common treatment effect and can handle irregularly spaced observations. Its correlation structure implies that measures on two different subjects are independent and measures on the same occasion between two variables are more highly correlated than measures taken on the same two variables but lagged over time. Although formulated for a single continuous and a single ordinal outcome the model extends to multiple binary, ordinal and continuous outcomes measuring the same underlying process.

3 Maximum Likelihood Estimation and Model Comparison

3.1 Adaptive Gaussian Quadrature

The marginal log-likelihood involves integration over the random effects distribution and over the error distribution of the latent process. The main challenge is that the integrals are nested and the inner integrand depends on the values of the random effects $\mathbf{b}_i$ which are unknown and have not been integrated out yet. To solve this problem Rabe-Hesketh, Skrondal and Pickles (2002) developed an iterative adaptive Gaussian quadrature procedure for estimation of multilevel models and implement the approach in gllamm in Stata. When the dimension of integration is large, an extension of the Monte Carlo Estimation Conditional maximization method of Gueorguieva and Agresti (2001) can be used.

TABLE 2. *Adaptive Gauss-Hermite quadrature estimates for the Depression Data Example.*

Parameter	HAMD only Estimate(SE)	CGI only Estimate(SE)	HAMD-CGI Estimate(SE)
Intercept	30.57(1.33)	–	30.47(1.32)
Factor loading	–	–	3.79(0.48)
Threshold 1	–	-6.60(0.61)	-7.14(0.87)
Threshold 2	–	-3.70(0.39)	-4.99(0.67)
Threshold 3	–	-1.71(0.30)	-2.83(0.49)
Threshold 4	–	0.92(0.30)	1.01(0.39)
Treatment	-1.60(1.74)	-0.25(0.42)	-0.40(0.49)
Time	-4.63(0.53)	-0.56(0.22)	-1.20(0.22)
Treatment×time	-1.28(0.92)	-0.48(0.31)	-0.35(0.25)
Time^2	0.33(0.07)	0.02(0.02)	0.10(0.02)
Treatment×time^2	0.11(0.13)	0.04(0.04)	0.03(0.04)
Var. of rand. intercept	30.60(7.76)	0.91(0.43)	2.09(0.72)
Var. of rand. slope	2.57(0.73)	0.26(0.09)	0.18(0.07)
Covar. of rand. effects	-3.61(1.87)	-0.09(0.13)	-0.22(0.13)
Var. of latent variable	–	–	0.95(0.18)
Var. of HAMD	15.04(1.48)	–	1.46(1.94)
Log-likelihood	-952.4	-284.5	-1154.84

3.2 Model Comparison

It is of interest to compare the fit of the joint model relative to the fit of separate models for the two response variables. However since the models are not nested the likelihood ratio test can not be applied. Here we use the ratio of geometric means of the contributions to the maximized likelihood $\rho_{m_1 m_2}$ proposed by Agresti and Caffo (2002). Values greater than 1 indicate that model m_1 is better than model m_2.

4 Results

Table 2 contains parameter estimates obtained using the two separate models for the two outcome measures and the joint model.
In all models all effects involving treatment are non-significant indicating that treatments are not significantly different. Due to the difference in scale the estimates from the continuous only model are not directly comparable to the estimates from the joint models. Since the estimate from the continuous only model is equal to $b = 3.79$ times the estimate from the joint model we can use the delta method to obtain directly comparable parameter estimates. In general, regression estimates from the joint models are in

between the estimates from the continuous only and the ordinal only models, and standard errors are similar. The likelihood based model comparison reveal that the joint model fits better than the separate models $\hat{\rho}_{JS} = 5.16$ and this model provides the best predictions at the subject level. A limited simulation study suggests that separate fitting of the continuous response variable is almost as efficient as joint fitting of the outcome variables and that in small samples the bias in regression parameter estimates may be smaller in the separate continuous model than in the joint model. On the other hand the separate model for the ordinal outcome shows larger bias and larger standard errors of the regression parameter estimates than the continuous only and the joint models.

5 Discussion

The results from the data example suggest that CGI does not contribute a significant amount of additional information about the latent process in excess of what is contained in HAMD and in view of the complexity of the joint model, a continuous only model may be preferrable. It is possible to extend the model to the case of two or more latent outcomes when each of these latent outcomes is measured via a different set of predictor variables. The model formulation can also be extended to situations with informative dropout by jointly modeling the outcomes and the probability of dropout.

References

Agresti, A. and Caffo, B. (2002). Measures of relative model fit. *Computational Statistics and Data Analysis*. **39**, 127–136.

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.

Rabe-Hesketh, S., Pickles, A., and Skrondal, A. GLLAMM: A class of models and a Stata program. (2001). *Multilevel Modelling Newsletter*, **13**, 17–23.

Rabe-Hesketh, S., Skrondal, A., and Pickles A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, **2**, 1–21.

Roy, J. and Lin, X. (2000). Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics*. **56**, 1047-1054.

Sanacora, G., Berman, R.M., et al. Yohimbine addition to SSRI therapy hastens the onset of antidepressant response. Submitted.

Multilevel Survival Analysis using Hierarchical Likelihood

Il Do Ha[1], Youngjo Lee[2], and Geon-Ho Cho[1]

[1] Faculty of Information Science, Kyungsan University, Kyungsan, 712-240, South Korea. Email: idha@kyungsan.ac.kr
[2] Department of Statistics, Seoul National University, Seoul, 151-742, South Korea. Email: youngjo@plaza.snu.ac.kr

Abstract: Nested survival data may be modelled by random-effect models such as multilevel mixed linear models and frailty models. We show how to analyze the well known chronic granulomatous disease data, which comprise recurrent infection times of patients from different hospitals, using both types of multilevel survival models and discuss their relative merits. Inference is based upon hierarchical likelihood, which provides a simple unified framework and a numerically efficient fitting algorithm for various random-effect models.

Keywords: Hierarchical likelihood; Multilevel frailty models; Multilevel mixed linear models; Nested survival data.

1 Introduction

Multilevel (or nested) structures often arise in biomedical research. We analyze a set of data on chronic granulomatous disease (CGD, Fleming and Harrington, 1991). Recurrent infection times of patients from different hospitals were observed. Both hospital and patient effects can be considered as random, with patient effect being nested within hospital. For analysis of survival data, there are two ways of introducing random effects. The frailty model (FM) introduces random effects into the hazard rate of recurrent event times, while the mixed linear model (MLM) introduces them into the expected values of recurrent event times. FMs specify the fixed- and random-effects multiplicatively on the conditional hazard rate, while MLMs specify them additively on the mean of recurrent event times. FMs are semi-parametric and fairly flexible, and their covariates can be time-dependent. However, FMs have been mainly developed for survival data analysis. MLMs have been widely used in many other areas, so that interpretation of their fixed and random effects is more familiar to statisticians. Because censored observations can be handled they can be useful alternatives to FMs for analysis of multivariate survival data. It is well known (Goldstein, 1995) that ignoring important sources of random variation may render traditional methods of statistical analysis invalid. Thus, multilevel

random-effect models are of interest. However, the difficulties encountered in extending these models prevent them from being well developed. Sastry (1997), Bolstad and Manda (2001) and Yau (2001) all studied limited classes of multilevel FMs as we shall see in Section 3. Multilevel models assume that random effects have a nested structure, which is not necessary for our approach. As far as we know, there is no existing literature on multilevel MLMs for analyzing nested survival data.
Recently Lee and Nelder (1996, 2001) have introduced the hierarchical likelihood (h-likelihood) which avoids the intractable integrals necessary to obtain the marginal likelihood. The h-likelihood gives a straightforward generalization of the fast and statistically efficient fitting algorithm for both single random-effect FMs and MLMs (Ha, Lee and Song, 2001, 2002) to their multilevel forms (nested and/or crossed).

2 The CGD Data

The CGD data set in Fleming and Harrington (1991) consists of a placebo-controlled randomized trial of gamma interferon (γ-IFN) in CGD. The aim of the trial was to investigate the effectiveness of the gamma interferon on serious infections in CGD patients. In this study, 128 patients from 13 hospitals were followed for about 1 year. The number of patients in a hospital ranges from 4 to 26. Of the 63 patients in the treatment group, 14 patients experienced at least one infection and a total of 20 infections were recorded. In the placebo group, 30 out of 65 patients experienced at least one infection, with a total of 56 infections being recorded.
Let T_{ijk} be the infection time for the kth observation of the jth patient in the ith hospital. In the CGD study about 63% of the data were censored. The recurrent infection times for a given patient are likely to be correlated. However, since each patient belongs to one of the 13 hospitals, the correlation may also be due to a random hospital effect. Yau (2001) developed multilevel log-normal FMs, in which infections, patients and hospitals are respectively defined as level 1, level 2 and level 3 units. He considered a single fixed covariate x_{ijk} (= 0 for placebo and = 1 for gamma interferon). The estimation of the variances of the random effects was also of interest. Throughout the paper, we let U_i be the unobserved frailty (or random effect) on the ith hospital and let U_{ij} be that on the jth patient in the ith hospital. We assume that the frailties U_i and U_{ij} are mutually independent and have density functions with frailty parameters α_1 and α_2, respectively.

3 Multilevel Frailty Models

Consider the multilevel FM below. Given $U_i = u_i$ and $U_{ij} = u_{ij}$, the conditional hazard function of T_{ijk} is of the form

$$\lambda_{ijk}(t|u_i, u_{ij}) = \lambda_0(t) \exp(x_{ijk}^T \beta) u_i u_{ij}, \tag{1}$$

TABLE 1. *Analyses using multilevel FMs.* $\alpha_1(\alpha_2)$*, the variance of the hospital (patient) frailty. M1, the Cox model* ($\alpha_1 = \alpha_2 = 0$)*; M2, two-level FM(*$\alpha_1 > 0$*,* $\alpha_2 = 0$*); M3, two-level FM(*$\alpha_1 = 0$*,* $\alpha_2 > 0$*); M4, three-level FM with both frailties.* $\widehat{\beta}_1$*, the estimate of* γ*-IFN effect* β_1*; ( ), the corresponding estimated standard error.* h_P*, the adjusted profile h-likelihood.*

Model	$\widehat{\beta}_1$	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$-2h_P$
M1	-1.086	—	—	707.42
	(0.268)			
M2	-1.119	0.157	—	703.62
	(0.269)			
M3	-1.062	—	0.778	693.24
	(0.321)			
M4	-1.067	0.024	0.750	693.20
	(0.319)			

where $\lambda_0(\cdot)$ is a nonparametric baseline hazard function, $\beta = (\beta_1, \ldots, \beta_p)^T$ is a $p \times 1$ vector of regression parameters and $x_{ijk} = (x_{ijk1}, \ldots, x_{ijkp})^T$ is a vector of fixed covariates. Following Yau (2001), we assume log Normal frailties; $V_i \equiv \log U_i \sim N(0, \alpha_1)$ and $V_{ij} \equiv \log U_{ij} \sim N(0, \alpha_2)$, so that the zero variances represent the absence of corresponding random components. For example, $\alpha_1 = \alpha_2 = 0$ corresponds to the Cox proportional hazards model. If $\alpha_1 = 0$ but $\alpha_2 > 0$, the model (1) becomes a two-level model without random hospital effects. Similarly, if $\alpha_2 = 0$ but $\alpha_1 > 0$, the model is without random patient effects. If both $\alpha_1 > 0$ and $\alpha_2 > 0$, the model (1) becomes a three-level model, requiring both random patient and random hospital effects. With FMs we can model the hazard rate of a series of infections in CGD patients.

The results are summarized in Table 1. For the three-level model our results are very similar to those of Yau (2001) ignoring ties. For example, in the three-level FM, we have $\widehat{\beta_1} = -1.067$ with SE (standard error) = 0.319, $\widehat{\alpha}_1 = 0.024$ and $\widehat{\alpha}_2 = 0.750$, while Yau (2001) has $\widehat{\beta_1} = -1.069$ with SE = 0.320, $\widehat{\alpha}_1 = 0.025$ and $\widehat{\alpha}_2 = 0.758$. When there are no ties Yau's method is identical to ours. However, the estimate of frailty parameters can be sensitive to ties (Therneau and Grambsch, 2000, pp. 250). Lee and Nelder (1996) showed that the deviance ($-2h_P$ in Tables 1 and 2) can be used for testing the absence of a random component. Here, h_P is the adjusted profile h-likelihood for dispersion components after eliminating nonparametric baseline cumulative hazards $\Lambda_0(t)$, fixed-effects β and random-effects v. Note that such a hypothesis is on the boundary of the parameter space, so the critical value is $\chi^2_{2\lambda}$ for a size λ test (Chernoff, 1954). For testing the absence of random-hospital effects $H_0 : \alpha_1 = 0, \alpha_2 > 0$ the deviance

difference is 0.04, which is not significant at a 5% level ($\chi^2_{1,0.10} = 2.71$), indicating that the random-hospital effects are not necessary. For testing the absence of random-patient effects $H_0 : \alpha_1 > 0, \alpha_2 = 0$ the deviance difference is 10.42, indicating that the random-patient effects are adequate. In addition, the deviance difference between the random-patient-effect only model ($\alpha_1 = 0, \alpha_2 > 0$) and the one-level Cox model ($\alpha_1 = \alpha_2 = 0$) is 14.18, i.e. the two-level model with the random-patient effect only appears to fit the data best, agreeing with Yau's (2001) conclusion. In the final lognormal FM, $\widehat{\beta_1} = -1.062$ (SE = 0.321) suggests that γ-IFN significantly reduces the rate of serious infections for CGD patients. With the h-likelihood we can use other frailty distributions. We tried gamma FMs, not reported here, and they provide similar results to the lognormal FMs. The choice of frailty distribution is relatively unimportant in inferences about fixed effects: see Sastry (1997) and Ha and Lee (2003) for FMs, and Ha et al (2002) for MLMs.

4 Multilevel Mixed Linear Models

For the responses $\log T_{ijk}$, we consider the three-level MLM

$$\log T_{ijk} = x_{ijk}^T \beta \ + \ U_i \ + U_{ij} \ + \ \epsilon_{ijk}, \tag{2}$$

where $x_{ijk} = (1, x_{ijk1}, \ldots, x_{ijkp})^T$ is a vector of fixed covariates, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is a $(p+1) \times 1$ vector of fixed effects, $U_i \sim N(0, \alpha_1)$, $U_{ij} \sim N(0, \alpha_2)$, and $\epsilon_{ijk} \sim N(0, \phi)$ are mutually independent and ϕ is the within-dispersion component. In MLMs the covariate x_{ijk}^T includes the intercept term, while in FMs the intercept term is not necessary, since it is confounded with the baseline hazard. With MLMs we directly model the recurrent times of serious infections in CGD patients. Note here that $(\alpha_1 = 0, \alpha_2 = 0)$ corresponds to one-level regression model without random effects, $(\alpha_1 = 0, \alpha_2 > 0)$ to two-level model without hospital effects, $(\alpha_2 = 0, \alpha_1 > 0)$ to two-level model without patient effects, and $(\alpha_1 > 0, \alpha_2 > 0)$ to three-level model, requiring both patient and hospital effects. Without censoring, model (2) is a standard MLM whose inferential procedures have been well developed. MLMs allowing censoring have been studied by a few authors, for example, Klein et al (1999), and Ha et al (2002). However, they have been restricted to single random-effect models. With the h-likelihood we can easily extend Ha et al's (2002) method to multilevel MLMs. Ha et al (2002) demonstrated by a numerical study that the h-likelihood procedure is robust against violations of the normal assumption.

The results are given in Table 2. As with analyses of the FMs we use the deviance $(-2h_P)$ for testing the absence of a random component. Here, h_P is the adjusted profile h-likelihood for dispersion components after eliminating fixed and random effects. The deviance $(-2h_P)$ shows again that the

TABLE 2. *Analyses using multilevel MLMs.* $\alpha_1(\alpha_2)$*, the variance of the random hospital(patient) effect;* ϕ*, within-variance component. M1, the regression model* ($\alpha_1 = \alpha_2 = 0$)*; M2, two-level MLM(*$\alpha_1 > 0$, $\alpha_2 = 0$*); M3, two-level MLM(*$\alpha_1 = 0$, $\alpha_2 > 0$*); M4, three-level MLM with both random effects.* $\widehat{\beta}_0(\widehat{\beta}_1)$*, the estimate of the intercept* β_0 *(*γ*-IFN effect* β_1*); ( ), the corresponding estimated standard error.* h_P*, the adjusted profile h-likelihood.*

Model	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\phi}$	$-2h_P$
M1	5.428	1.494	–	–	3.160	426.52
	(0.185)	(0.322)				
M2	5.594	1.470	0.294	–	2.872	422.00
	(0.249)	(0.313)				
M3	5.661	1.237	–	0.722	2.163	417.60
	(0.202)	(0.331)				
M4	5.684	1.262	0.085	0.635	2.182	417.24
	(0.220)	(0.328)				

two-level MLM with only the random patient effect fits the data best. In the final MLM $\widehat{\beta_1} = 1.237$ (SE = 0.331) means that the γ-IFN significantly prolongs the recurrent infection times.

5 Discussion

In the CGD data the proportional hazards assumption may be suspect (Lindsey, 1995). By introducing frailties in FMs we may overcome such a restriction (Keiding et al, 1997). Indeed the deviance test shows that we need random effects and the two-level models with only random patient effects are chosen as final models among models we considered. FMs and MLMs lead to equivalent conclusions; FMs show that γ-IFN reduces the hazard rate for serious infections, while MLMs show that it prolongs the recurrent infection times. We concluded via a residual analysis (not shown) that both models are equally plausible. Because both models fit the data equally well, we may use a FM when interest is on reduction of the hazard ratio or a MLM when it is on prolonging of recurrent infection times.

References

Bolstad, W.M. and Manda, S.O.M. (2001). Investigating child mortality in Malawi using family and community random effects: A Bayesian analysis. *Journal of the American Statistical Association*, **96**, 12–19.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics*, **25**, 573-578.

Fleming, T.R. and Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. New York: Wiley.

Goldstein, H. (1995). *Multilevel Statistical Models*. London: Arnold.

Ha, I.D. and Lee, Y. (2003). Estimating frailty models via Poisson hierarchical generalized linear models. *Journal of Computational and Graphical Statistics*. (in press).

Ha, I.D., Lee, Y., and Song, J.-K. (2001). Hierarchical likelihood approach for frailty models. *Biometrika*, **88**, 233–243.

Ha, I.D., Lee, Y., and Song, J.-K. (2002). Hierarchical likelihood approach for mixed linear models with censored data. *Lifetime Data Analysis*, **8**, 163–176.

Keiding, N., Andersen, P.K., and Klein, J.P. (1997). The role of frailty models and accelerated failure time models in describing heterogeneity omitted covariates. *Statistics in Medicine*, **16**, 215–224.

Klein, J.P., Pelz, C., and Zhang, M. (1999). Modelling random effects for censored data by a multivariate normal regression model. *Biometrics*, **55**, 497–506.

Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B*, **58**, 619–678.

Lee, Y. and Nelder, J.A. (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, **88**, 987–1006.

Lindsey, J.K. (1995). Fitting parametric counting processes by using log-linear models. *Journal of the Royal Statistical Society, Series C*, **44**, 201–212.

Sastry, N. (1997). A nested frailty model for survival data, with application to study of child survival in Northeast Brazil. *Journal of the American Statistical Association*, **92**, 426–435.

Therneau, T.M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer-Verlag.

Yau, K.K.W. (2001). Multilevel models for survival analysis with random effects. *Biometrics*, **57**, 96–102.

The Behavior of the Likelihood Ratio Test for Testing Missingness

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

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

Abstract: To asses the sensitivity of conclusions to model choices in the context of selection models for non-random dropout, one can oppose the different missing mechanisms to each other; e.g. by the likelihood ratio tests. The finite sample behavior of the null distribution and the power of the likelihood ratio test is studied under a variety of missingness mechanisms.

Keywords: Missing data; Sensitivity analysis; Likelihood ratio test; Missing mechanisms.

1 Introduction

In a longitudinal setting, units are measured on several occasions. It is not unusual that a sequence is not fully observed, due to intermediate missingness and dropout. In the context of maximum likelihood inference, Rubin (1976) classified missing data into three types, namely missing completely at random, missing at random and missing not at random. Diggle and Kenward (1994) use a selection model to represent such a process. A selection model consists of two parts: a measurement part and a missingness process part. Such a model relies on strong and untestable assumptions. Not only the distributional assumptions can be misspecified but also the presence of missing data can have a large impact. In classical theory, the asymptotic distribution of the likelihood ratio test is a chisquare distribution with degrees of freedom equal to the difference in number of parameters. Careful considerations have to be made when using this result to test for missing not at random as shown by Rotnitzky et al (2000). We will first provide a motivating example from Rotnitzky et al (2000), then we will introduce selection models. In a simulation study, we will illustrate the finite sample behavior of the likelihood ratio test and we will conclude with some current research topics.

2 A Motivating Example

The following example is used in Rotnitzky et al (2000). Let $Y_1, \cdots, Y_n$ be a sample of n observations from a normal distribution with mean β and

variance σ^2. Suppose there is missingness in this sample which is possibly related to the outcome itself. Let us denote this conditional probability by

$$P_c(y;\alpha_0,\alpha_1) = e^{H(\alpha_0+\alpha_1(y-\beta)/\sigma)}$$

where α_0 and α_1 are unknown parameters and $H(.)$ is a known function assumed to have its first three derivatives at α_0 non-zero. Interest goes out to test whether $\alpha_1 = 0$ which corresponds to missing completely at random. We thus consider two random variables (R, Y) where R is a binary indicator, which is 1 if Y is observed and 0 otherwise. The contribution of one individual to the loglikelihood is thus

$$\begin{aligned} & r[-\log\sigma - (y-\beta)^2/(2\sigma^2) + H\{\alpha_0 + \alpha_1(y-\beta)/\sigma\}] \\ + \quad & (1-r)[\log E\{1 - P_c(y;\alpha_0,\alpha_1)\}] \end{aligned}$$

For n individuals the loglikelihood $L_n(\beta,\sigma,\alpha_0,\alpha_1)$ is the sum of n such terms., If we have a look at the score vector at the null point $\beta,\sigma,\alpha_0,\alpha_1 = 0$ we obtain the following equations.

$$\begin{aligned} & r(y-\beta)/\sigma^2 \\ & r(-\sigma^2 + (y-\beta)^2)/\sigma^3 \\ & rH'(\alpha_0) - (1-r)\frac{H'(\alpha_0)e^{H(\alpha_0)}}{1-e^{H(\alpha_0)}} \\ & rH'(\alpha_0)(y-\beta)/\sigma \end{aligned}$$

We can see that this score vector is degenerate at this particular parameter point. Equivalently, the information matrix calculated from expected second derivatives is singular at this parameter point.
Rotnitzky et al (2000) show that likelihood-based inference with a singular information matrix can have some consequences with respect to the distribution of the likelihood ratio test. Depending on the nature of the model either the asymptotic distribution can be a mixture of χ^2-distributions or the convergence rate is very slowly. Due to these demerits the application of the asymptotic distribution has to be considered with care. We will illustrate this behavior in the context of selection models by simulations.

3 Selection Models

Let us assume that for subject i, $i = 1,\cdots,N$, a sequence of responses Y_{ij} is measured at several occasions $j = 1,2,\ldots,J$. Let R_{ij} be a missingness indicator and assume that y_{i1} is always observed. Then $r_{ij} = 0$ if y_{ij} is missing and $r_{ij} = 1$ if y_{ij} is observed. The measurement part of the model of Diggle and Kenward (1994) is given by

$$\mathbf{Y_i} = (Y_{i1},\ldots,Y_{iJ}) \sim N(X_i\beta, \Sigma_i), \qquad i = 1,\ldots,N,$$

where β is a vector of fixed effects, X_i is a matrix containing covariate values and Σ_i is a covariance matrix. The missingness process is described by

$$\text{logit}[Pr(R_{ij} = 1|y_{i,j-1}, y_{ij})] = \psi_0 + \psi_1 y_{i,j-1} + \psi_2 y_{ij},$$

where $Pr(R_{ij} = 1|y_{i,j-1}, y_{ij})$ is the probability for the i^{th} subject to drop out at time j. If ψ_2 differs from zero, the missingness process is non-random. Let us denote

$$g(\mathbf{h_{id}}, y_{id}) = Pr(R_{id} = 1|y_{i,d-1}, y_{id})$$

with d the time of dropout and $\mathbf{h_{id}} = (y_{i1}, \ldots, y_{i,d-1})$ the history of y_{id}, which we now restrict to depend on the previous measurement only. The total loglikelihood has the form

$$\ell = \sum_{i=1}^{N} (r_i \ell_i^c + (1 - r_i)\ell_i^i),$$

with ℓ_i^i the contribution for an incompleter

$$\ell_i^i = \ln f(\mathbf{h_{id}}) + \sum_{j=2}^{d_i-1} \ln[1 - g(\mathbf{h_{ij}}, y_{ij})] + \ln \int f(y_{id}|\mathbf{h_{id}}) g(\mathbf{h_{id}}, y_{id})\ dy_{id}$$

and ℓ_i^c the contribution for a completer

$$\ell_i^c = \ln f(\mathbf{y_i}) + \sum_{j=2}^{J} \ln[1 - g(\mathbf{h_{ij}}, y_{ij})].$$

The likelihood ratio test statistic for testing MNAR versus MAR is then given by

$$G = -2[\ell_{MNAR} - \ell_{MAR}].$$

Due to the difference in only one parameter, the distribution of this statistic can be misleadingly expected to be $\chi^2(1)$. Based on this statistic Kenward (1998) and Molenberghs et al (2001) rejected the null hypothesis of missing at random on a value of 5.11, which corresponds to a P-value of 0.02 for their data example (Mastitis in dairy cattle). They compared this result with the Wald test (P-value of 0.002) and concluded that the asymptotic approximations are not very accurate. Rotnitzky et al. (2000) state that the regular assumptions of the likelihood ratio test statistic do not hold in this case due to the singular information matrix. In the next paragraph, we will illustrate the behavior of the likelihood ratio test statistic for the different missingness parameters in a simple setting.

4 Simulations

For this small simulation study 400 similar datasets were generated in 4 different settings. Each dataset consists of 200 subjects, each with two measurements generated from a bivariate normal distribution. Consider the following bivariate normal distribution, based on a compound symmetry covariance matrix:

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 4 \\ 2 \end{pmatrix}, \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}\right]. \tag{1}$$

The dropout process was generated according to the following model

$$\text{logit}[P(R_i = 1|Y_{i1}, Y_{i2})] = -2 + \psi_1 Y_{i1} + \psi_2 Y_{i2} \tag{2}$$

where ψ_1 and ψ_2 were chosen according to four different settings. In setting 1, the null hypothesis is $\psi_1 = 0$, given that $\psi_2 = 0$, while in setting 2 the null hypothesis is $\psi_1 = 0$, given that $\psi_2 \neq 0$. Setting 3 considers a test for $\psi_2 = 0$, given that $\psi_1 = 0$ and finally in setting 4 $\psi_2 = 0$ is tested, given that $\psi_1 \neq 0$. In the next table an overview of the different simulation settings is given.

	Data under H_0 with	
	$\psi_2 = 0$	$\psi_2 \neq 0$
$H_0 : \psi_1 = 0$	Setting 1	Setting 2
	$\psi_1 = 0$	$\psi_1 \neq 0$
$H_0 : \psi_2 = 0$	Setting 3	Setting 4

Figure 1 shows plots of the simulated null-distributions together with approximating χ^2-distribution.

5 Discussion and Further Research

From the literature and the simulation settings, it is clear that the likelihood ratio test for testing missing not at random does not fulfill the regular assumptions. The use of classical asymptotic results might clearly lead to false results. A study of the theoretical asymptotical distribution and a power simulation study are topics of current research.

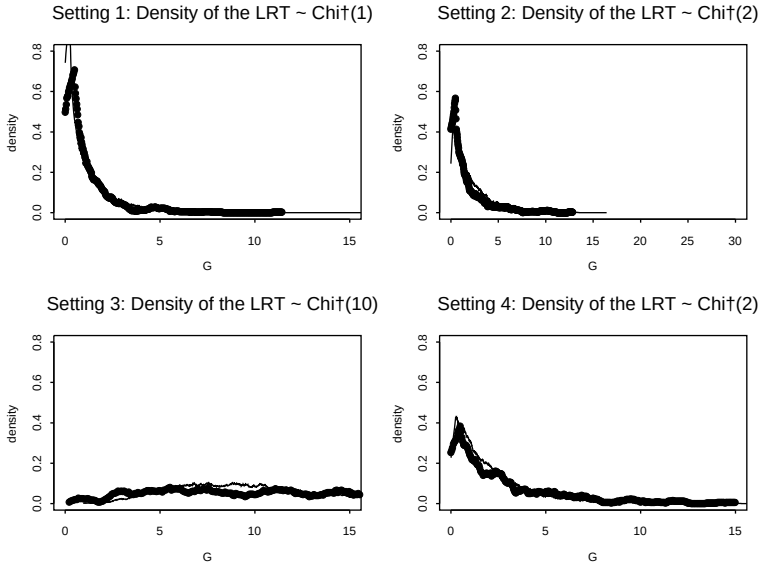


FIGURE 1. *Density plots (dots) of the different settings with approximating* χ^2*-distribution (full line).*

References

Diggle, P.J. and Kenward, M.G. (1994). Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics*, **43**, 49–93.

Kenward, M.G. (1998). Selection models for repeated measurements with nonrandom dropout: An illustration of sensitivity. *Statistics in Medicine*, **17**, 2723-2732.

Little, R.J.A. (1976). Inference about means from incomplete multivariate data. *Biometrika*, **63**, 593–604.

Molenberghs, G., Verbeke, G., Thijs, T., Lesaffre, E., and Kenward, M.G. (2001). Influence analysis to assess sensitivity of the dropout process. *Computational Statistics and Data Analysis*, **37**, 93–113.

Rotnitzky A., Cox D.R., Bottai M., and Robins J. (2000). Likelihood-based inference with singular information matrix. *Bernouilli*, **6**(2), 243–284.

Using Non-Parametric Estimators to Model a Monotonic Dose Response Curve and Bootstrap Confidence Intervals

Ian Hirsch[1]

[1] Clinical Statistics Europe, Pfizer Global Research and Development, Ramsgate Rd, Sandwich, UK, CT13 9NJ

Abstract: In this paper we consider study designs which include a placebo and an active control group as well as several dose groups of a new drug. A monotonically increasing dose response function is assumed, and the objective is to estimate a dose with equivalent response as the active control group, including a confidence interval for this dose.
We present different non-parametric methods to estimate the monotonic dose response curve. One is based upon the well known isotonic regression estimator, and the other one upon a non-negative least squares estimator. We introduce a bias correction to overcome a bias for the second method.
We also use two different bootstrap methods to obtain the confidence intervals. One is based upon the standard bootstrap. The other method is slightly more sophisticated, and ensures that the resampling distributions comply with the order restrictions imposed.
In our simulations we did not find any differences between the two bootstrap methods. The non-negative least squares estimator yields biased results for moderate sample sizes. The bias adjustment for this estimator works well, even for small and moderate sample sizes. Surprisingly, we also found that this bias adjusted non-negative least squares method outperforms the isotonic regression method in some situations, but we did not find any situations where the isotonic regression method performs better. (Dilleen et al (2003))

Keywords: Monotonic dose response; Isotonic regression estimator; Restricted least squares estimator; Bootstrap confidence intervals.

1 Introduction

The statistical modelling of dose-response relationships is a common problem in the pharmaceutical industry, which occurs in quantal bioassays, toxicology experiments, and clinical dose finding studies, as well as in many other situations. In this presentation we are interested in a clinical application of dose-response analysis, and in the estimation of a dose which has an equivalent effect to the active comparator group.
We consider a study design which includes a placebo group ($d_0 = 0$), several dose groups ($d_1, d_2, \ldots d_I$) of a new compound, and an active control group,

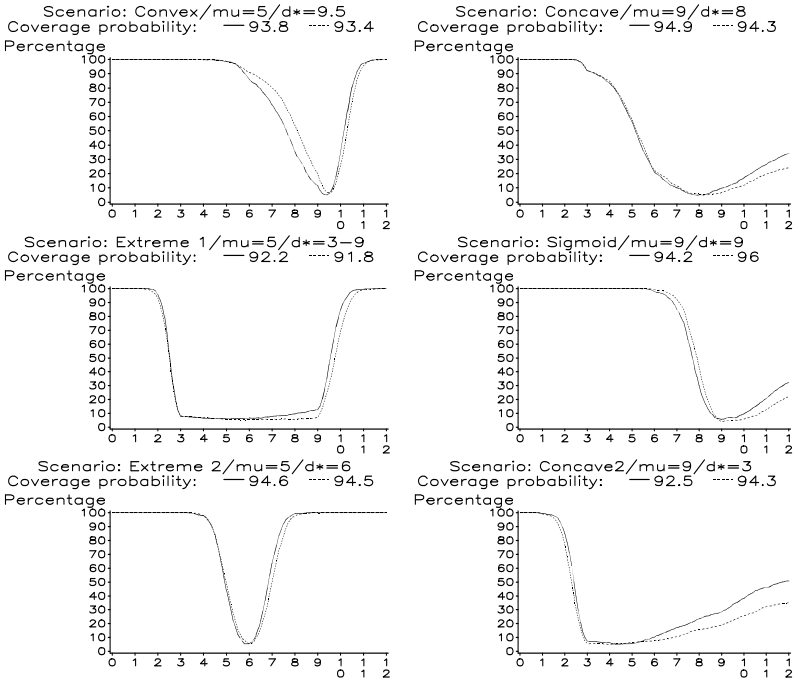


FIGURE 1. *Comparison of Confidence Intervals Based Upon the Isotonic Regression Estimator and the Bias-Adjusted Non-Negative LSE. The bold lines refer to the bias-adjusted non-negative lse, and the dotted lines refer to isotonic regression estimator. The number of patients per group was* $n_i = 60$*, and the standard deviation was* 5*.) (see Dilleen et al, 2003)*

which is usually the standard treatment for the respective disease. We assume there are n_0 patients in the placebo group, n_i patients in the dose groups ($i = 1, \ldots, I$), and n_a patients in the active control group. The endpoint of interest is continuous, and we denote the observations in the placebo group and in the dose groups as $Y_{i1}, \ldots, Y_{in_i}$ (for $i = 0, \ldots, I$). $Y_{a1}, \ldots, Y_{an_a}$ are the observations in the active control group. The observed mean values of the corresponding groups are denoted as $\bar{Y}_{i\cdot}$ or $\bar{Y}_{a\cdot}$. The expected means $E[Y_{i\cdot}] = f(d_i)$ are assumed to be monotonic.

2 Estimating the Dose Response Curve and the Confidence Intervals for the Equivalent Dose

We use different methods to model the monotonic dose response curve. These methods have one feature in common: based upon the observed

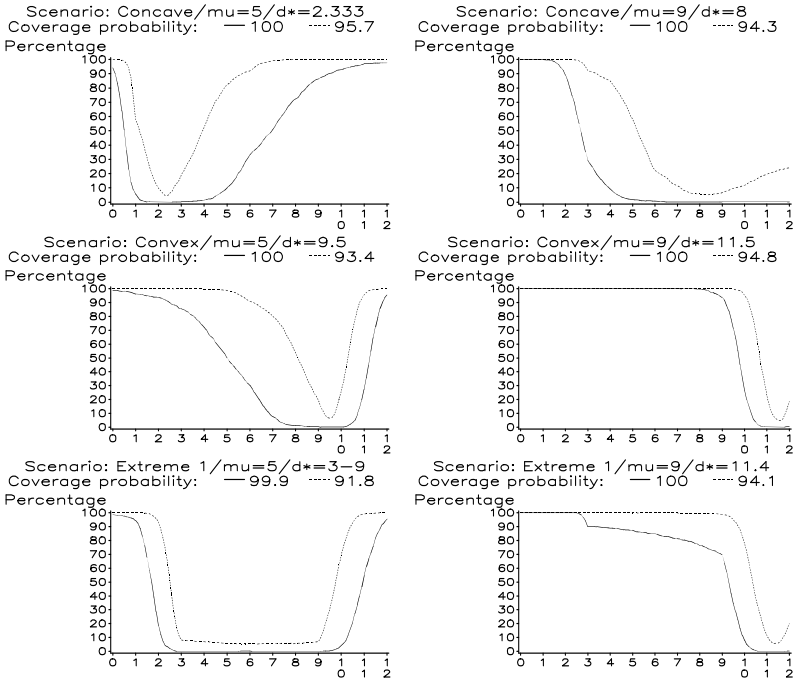


FIGURE 2. *Comparison of Confidence Intervals Based Upon the Bootstrap and Based Upon Korn's Method using the Isotonic Regression Estimator (The dotted lines refer to the bootstrap confidence band, and the bold lines refer to Korn's method using the isotonic regression estimator. The number of patients per group was* $n_i = 60$*, and the standard deviation was* 5*.) (see Dilleen et al 2003)*

means $\bar{Y}_{0\cdot}$ to $\bar{Y}_{I\cdot}$ a monotonic sequence $\hat{f}(d_0) \leq \hat{f}(d_1) \leq \ldots \leq \hat{f}(d_I)$ will be estimated. The monotonic dose response function $\hat{f}(d)$ is then obtained by linear interpolation between $\hat{f}(d_i)$ and $\hat{f}(d_{i+1})$. The dose $\hat{d}$ which provides an equivalent response as the active control is the solution to $\bar{Y}_{a\cdot} = \hat{f}(\hat{d})$. A confidence band around this estimated dose $\hat{d}$ is obtained through bootstrap techniques.

The first method is based upon the **isotonic regression estimator**. By linearly interpolating between these monotonic estimates both the estimated dose $\hat{d}$ and a bootstrapped 95% confidence interval around this dose can be calculated. Simultaneous confidence bands for $f(d)$ are also found theoretically using the methods described in Korn (1982). From these, theoretical confidence intervals around $\hat{d}$ are found and compared with the bootstrap method.

TABLE 1. *Mean Values for Different Dose Response Scenarios. The table displays the mean values $f(d_0), f(d_1), \ldots$ for six different dose response scenarios. The underlying study includes a placebo control and five doses (1mg, 3mg, 6mg, 9mg, and 12mg) of a new drug.*

scenario	$d_0 = 0$	$d_1 = 1$	$d_2 = 3$	$d_3 = 6$	$d_4 = 9$	$d_5 = 12$
convex	0	0.2	0.5	2	4	10
concave	0	3	6	8	9.5	10
sigmoid	0	0.25	1	5	9	10
linear	0	0.83	2.5	5	7.5	10
anti-sigm.	0	1.4	4	5	6	10
extreme 1	0	0	5	5	5	10
extreme 2	0	0	0	5	10	10

The second method is based upon the new **non-negative least squares estimator** and described in Dilleen et al (2003). Here we assume that the dose response curve $f(d)$ is again monotonic, i.e. $f(d_0) \leq f(d_1) \leq \ldots \leq f(d_I)$. We assume that the observations $y_{ij} = f(d_i) + \varepsilon_{ij}$, and the residuals ε_{ij} are generated from the same distribution and are all independent observations with zero mean and the same variance. The dose response functions are assumed to be made up of of a linear combination of monotonic base functions $g_i(d)$ such that $f(d) = \alpha_0 + \sum_{i=1}^{I} \beta_i g_i(d)$. The parameters β_i are initially estimated by the ordinary least squares estimator and in order to comply with the order restrictions, the least squares estimator is modified accordingly where the $\hat{\beta}_i^+$ are equated to zero if negative. We also assume that the base functions are the distribution functions of uniform distributions on $[d_{i-1}, d_i]$ which result in $f(d)$ being a piecewise linear monotonic function between $f(d_{i-1})$ and $f(d_i)$. By modelling the dose response curve using these monotonic estimates the dose which is equivalent to the active control and its 95% confidence interval is obtained as they were for the isotonic regression estimates.
However, it is easily seen via our extensive simulations that this method is biased and therefore a third method based upon a **bias adjusted non-negative least squares estimator** has been developed using a suitable bias adjustment to the non-negative least squares estimate (see Dilleen et al, 2003).

3 Comparison of the Methods Using Simulations

Main findings from extensive simulations, carried out using different underlying scenarios based on a study design which includes 5 doses of a new compound together with placebo and a competitor drug will be presented.

We conducted an extensive simulation study to compare the three estimation methods as well as to compare the different methods to obtain confidence bands (see Dilleen et al 2003). Different underlying scenarios were used for these simulations. They are based upon a dose finding study with $I = 5$ dose groups ($d_1 = 1mg, d_2 = 3mg, d_3 = 6mg, d_4 = 9mg$, and $d_5 = 12mg$). These scenarios are explained in Table 1, where the mean values of the dose groups (including the placebo group $d_0 = 0$) are displayed. The active control group means μ were also varied and for each scenario and each level of μ.
These simulations demonstrate that, surprisingly the new bias-adjusted least square method outperforms the well-known isotonic regression method in certain situations (see Figure 1). However there are not many situations where the isotonic regression method performs better. Also using bootstrap confidence intervals are shown to clearly outperform the corresponding theoretical approach (see Figure 2).

References

Dilleen, M., Heimann, G., and Hirsch, I. (2003). Non-parametric estimators of a monotonic dose response curve and bootstrap confidence intervals. *Statistics in Medicine*, **22**, 869-882.

Korn, E.L. (1982). Confidence bands for isotonic dose response curves. *Applied Statistics*, **31**, 59-63.

Investigation into Drivers for Flowering in Eucalypts: Effects of Climate on Flowering

Irene L. Hudson[1], Adrian Barnett[2], Marie R. Keatley[3], and Peter K. Ades[3]

[1] Department of Mathematics and Statistics, Canterbury University, Christchurch, New Zealand
[2] School of Population Health, University of Queensland, Queensland, Australia.
[3] School of Resource Management, University of Melbourne, Parkville, Australia

Abstract: Regardless of cyclicity of flowering over time, this study shows that the flowering intensity of *E leucoxylon* is significantly influenced by temperature and that the effect of temperature is non-linear. Upper and lower thresholds of flowering temperature for *E leucoxylon* have been confirmed and estimates of the long and short-term, non-linear effects of climate given.

Keywords: Bayesian adaptation of penalized regression splines; Generalised additive models; Discrete time series; Climate change.

1 Introduction

As reported recently in Keatley et al (2002), phenological study involves the recording of recurring natural events such as the commencement of flowering (Koch, 2000) or the arrival of migratory birds (Sparks, 1999), and the influence on such events by edaphic and climatic factors. Recently analyses of phenological data have been used to examine potential impacts of climate change and the observed global increase in temperature (Sagarin and Micheli, 2001). These studies, however, have thus far used data concentrated in the Northern Hemisphere. Currently, there are limited phenological studies undertaken at a research level in Australia (Keatley et al, 1999; Manning and Nobre, 2001). This study is one of the first attempts to utilise Australian phenological data to detect responses to climate change (Keatley et al, 2002). The data represents a long time series, for Australasian standards, using more than 30 years of monthly readings, in excess of 400 flowering/climate time points. This paper focuses on one species, *E. leucoxylon* (*E.l.*), but is part of a larger study examining eight Eucalypt species flowering profiles from Jan 1938-Mar 1972 (Keatley et al, 2002, 2000, 1999). The primary aim of this paper is to investigate the relationship between flowering intensity and temperature, alone or in combination with rainfall, since temperature is a major climatic influence on phenological events such as flowering (Snyder, 2001).

2 Statistical Methods

Method I: Generalised additive models (GAMs)
GAMs (Hastie and Tibshirani, 1990) were used to investigate the effect of rain and mean diurnal temperature simultaneously. In addition, the effect of a long-term time/year (TREND) was factored in. GAM models allow for non-linear relationships between the observed monthly discrete flower counts (range 0.0 - 5.0) and climatic predictors and trend, using smoothed regression lines. To estimate the non-linear regression lines, a local scoring algorithm and a spline with four degrees of freedom (Hastie and Tibshirani, 1990) was used. The results were generated using S-Plus (Venables and Ripley, 1997).
Method II: Bayesian adaptation of Penalized Regression Splines
An alternative, non-linear framework within a GAM model, is to estimate the regression lines using a penalized regression spline (Wood, 2000). The penalized regression spline (PRS) method was reformulated into a *Bayesian hierarchical framework* and the variance of the spline terms was used as the constraint, rather than as a constant smoothing parameter. The Bayesian adaption of the penalized regression spline (BAPRS) essentially estimates the degree and shape of the non-linearity. The Bayesian framework has the further advantage that missing covariates are easily dealt with, although in this data set the number of missing covariates was small. There were three values in total missing from monthly temperature from a total sample of 376 (0.8%). The Bayesian model was implemented using the WinBUGS package (Spiegelhalter et al, 1999). In total 30,000 MCMC simulations were run after a burn-in of 5,000.

3 Results

Method I: GAM models
The GAM model is:
$\text{Count}(t) = \text{Count}(t-1) + f(\text{MEANTEMP}) + f(\text{Rainfall}) + f(\text{TREND})$,
where f is an unspecified nonparametric function based on spline smoothers and a Poisson link function is used. The previous count (count at $t-1$) is used in the model as an autoregressive (AR(1)) term for the correlated counts time series. The focus is to model the effect of temperature and rainfall together, with the previous count treated as a nuisance parameter. GAM fits showed that mean diurnal temperature (MEANTEMP) is statistically significant ($p = 0.0228$), as is the previous count (LAGLEUCO). Rainfall (RAIN) and TREND are not statistically significant ($p > 0.5$). A GAM plot is given in Figure 1, where the y-axis represents the estimated change in flower count, which is the change from the mean flower count value versus mean diurnal temperature (TEMP) in °C. For example at the highest temperature (~ 24°C) the expected number of flowers would be

approximately one less than the mean. From the GAM modelling (Fig. 1), the estimated effects of temperature appear smoothly non-linear.
We note, that the observed points of shift in the level of the GAM TREND, at 20, 23/24 and 30 years, coincide with 1960, 1963/64 and 1970, when a significant ($p < 0.005$) shift in flowering was found by block bootstrap and survival change point (CP) analyses (Dalrymple et al, 2001). Further CP analyses also confirmed that 4 to 5 years prior to these times, temperature and rainfall had changed significantly ($p < 0.005$) in 1955, 1958/59 and 1965/6.

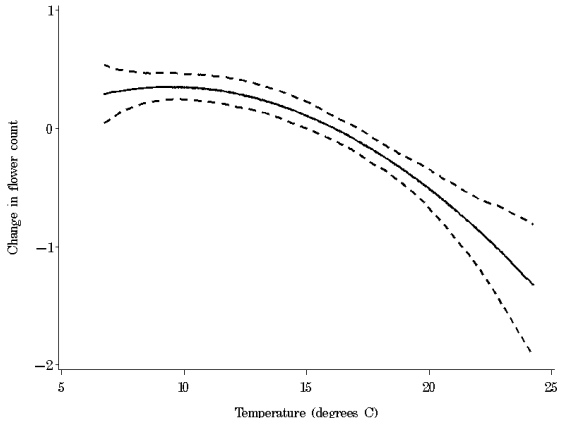


FIGURE 1. *Estimated effect of mean diurnal temperature and GAM 95% confidence interval (CI)*

Method II: Bayesian adaptation of Penalized Regression Splines
The effect of temperature is modelled, assuming that: monthly flower counts have a Poisson distribution, mean monthly temperature has a Normal distribution, and that the AR term for previous flower count is uniform and stationary (within -1 and 1). For the non-linear spline five knots were placed at 7.5, 12, 15, 17.5, and 22.5 degrees centigrade, and the precision of the linear spline terms restricted using a Gamma(4000, 20) prior distribution. Initial BAPRS models, as in the GAMs, rejected the significance of rainfall or trend on flower counts and these covariates were subsequently omitted from the final model. The effect of temperature is non-linear and the results using GAMs and BAPRS agree closely. In agreement with Figure 1 flower intensity below 10 to 12°C appears stable and significantly above the average. This agrees with work by Keatley et al (1999), which found that the lower threshold temperature for *E.l.* to induce flowering is 9.9°C. Bayesian posterior intervals, not reported here, and the GAM 95% confidence interval (Fig. 1) do not contain zero after 18°C, as at these temperatures there is a negative effect on flowering. This implies that an upper threshold tem-

perature for *E.l.* may be 18°C, not 25°C, as based on earlier work (Keatley et al, 1999).
The estimated correlation between neighbouring months was 0.41 with a 95% posterior interval of [0.33, 0.47]. The adapted Bayesian PRS predicted values and the observed counts match each other closely (Figure 2). The mean square error was 0.59, for 376 observations. According to the Kolmogorov-Smirnov test the model residuals were not normal due to the presence of three outliers. However, the residuals appear to be linearly uncorrelated according to the periodogram based test (Fuller, 1996).

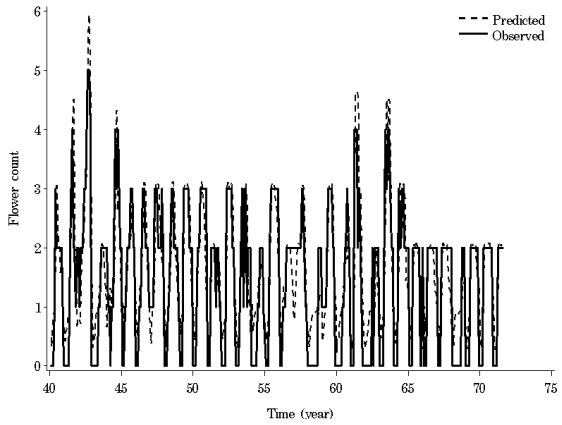


FIGURE 2. *Predicted and observed counts over time for the Bayesian penalized spline model*

4 Conclusion

Upper and lower thresholds of flowering temperature have now been identified for *E leucoxylon*. Keatley and Hudson (1998) showed there is an optimal time for a species to commence or cease flowering, depending on bud and fruit volume. Reproductive success may thus be influenced by shifts in flowering commencement. Changes in temperature are likely to translate to changes in the timing of *E leucoxylon* flowering commencement. Phenological indicators, such as flowering, may thus prove to be valuable proxy indicators of global climate change. This study shows that after accounting for temperature, the long-term trend in the number of flowers is relatively stable after 1944. Recent research by the authors, using spectral analysis, as in Legendre and Legendre, (1998), has shown that flowering intensity in *E leucoxylon* has a 2 year cycle. These results imply that there may be an internal mechanism, e.g., levels of nitrogen etc., that may also underpin flowering, once the effects of external environment are accounted for.

References

Dalrymple, M.L., Hudson, I.L., and Barnett, A.G. (2001). Survival, block bootstrap and mixture methods for detecting change points in discrete time series data with application to SIDS. *Proceedings of 16th IWSM*, 135–146.

Fuller, W. (1996). *Introduction to Statistical Time Series*. New York: Wiley.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*. London: Chapman & Hall.

Keatley, M.R, Fletcher, T.F, Hudson, I.L., and Ades, P.K. (2002). Phenological studies in Australia: Potential application in historical and future climate analysis. *International Journal of Climatology*, **22**(14), 1769–1780.

Keatley, M.R. and Hudson, I.L. (2000). Influences on the flowering phenology of three eucalypts. In *Biometeorology and Urban Climatology at the Turn of the Century.*, de Dear RJ, Kalma JD, Oke TR, Aucliems A (eds). World Meteorological Organisation: Geneva, Switzerland; 191–196.

Keatley, M.R., Hudson, I.L., and Fletcher, T.D. (1999). The use of long-term records for describing flowering behaviour: A case-study in Victorian Box-Ironbark Forests. In *Australia's Ever-changing Forests* IV, Dargavel J, Wasser B (eds). Australian University Press: Canberra; 311–328.

Keatley, M.R. and Hudson, I.L. (1998). The influence of fruit and bud volumes on Eucalypt flowering: an exploratory analysis. *Australian Journal of Botany*, **42**, 281–304.

Koch, E. (2000). Phenology in Austria: Phenological mapping - long-term trends. In *Biometeorology and Urban Climatology at the Turn of the Century.*, de Dear RJ, Kalma JD, Oke TR, Aucliems A (eds). World Meteorological Organisation: Geneva, Switzerland, 187–190.

Legendre, P. and Legendre, L. (1998). Numerical Ecology. Developments in Environmental modelling 20. Elsevier: Quebec Canada. pp 679–691.

Manning, M. and Nobre, C. (eds). (2001). Technical Summary Climate Change 2001: Impacts, adaptation, and vulnerability. *Intergovernmental Panel on Climate Change*. Geneva: Chapman & Hall/CRC.

Sagarin, R. and Micheli, F. (2001). Climate change in non-traditional data sets. *Science*, **294**, 811.

Snyder, R.L., Spano, D., Duce, P., and Cesaraccio, C. (2001). Temperature for phenological models. *International Journal of Biometeorology*, **45**, 178–183.

Sparks, TH. (1999). Phenology and the changing pattern of bird migration in Britain. *International Journal of Biometeorology*, **42**, 134–138.

Spiegelhalter D. J., Thomas, A., and Best, N.G. (1999). *WinBUGS Version 1.2 User Manual.* MRC Biostatistics Unit.

Venables, W.N. and Ripley, B.D. (1997). *Modern Applied Statistics with S-PLUS.* New York: Springer-Verlag.

Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society, Series B*, **62**(2), 413–428.

Issues and Trends in Modelling Internet Congestion

S. Ravi Jagannathan[1] and Kenan M. Matawie[1]

[1] School of Quantitative Methods and Mathematical Sciences, University of Western Sydney, Po Box 10, Kingswood NSW 2747, Australia.
Email: k.matawie@uws.edu.au, r.jagannathan@uws.edu.au

Abstract: This paper is intended to provide a theoretical treatment of the ways and means by which congestion in the Internet can be understood, modelled and subsequently forestalled. 3 popular schemes are presented and critiqued and our new model outlined. Pros and cons of these models are discussed.

Keywords: Congestion; TCP/IP (Transmission Control Protocol/Internet Protocol) protocol stack; TCP connections (flows), Congestion windows, Routers.

1 Introduction

This collection of prior results and techniques in the modeling of Internet congestion constitutes an overview of the subject that is not readily found anywhere else. The level of mathematical/statistical sophistication is deliberately kept minimal, the focus being on the introduction of key concepts, in a manner that detailed statistical analysis can be then carried out in a second stage. The fundamental scenario is that "sources" (ingress) send "packets" (sets of bytes of information) across the network, to "destinations" (egress). The network itself is a mesh of interconnected routers (switching/routing elements) and subnets. A basic assumption is that the network has an innate maximum "capacity". If too many packets enter the network, the network will discard some. This situation of "loss under overload" is called congestion. In this paper, we place the problem in perspective using the key notions of Black Box and White Box models. Then we sample survey a few now classical Black Box models, i.e., TCP Tahoe, TCP Reno, and TCP Vegas. We critique these three models, and present some of their relative merits and demerits. Our new contribution to Black Box modeling, S-Channels, is introduced briefly, in terms of its characteristics, properties and benefits.

2 Black Box & White Box Models

In this paper it is argued that it makes sense to differentiate between two approaches to congestion management, i.e., Black Box models and White

Box models. Black Box: Assume that the internal "structure" or innards of the network, in terms of its mesh of routers, is unknown. Howsoever routers manage their internal queues and buffers, given any such constellation of routers, how does one maximize end to end usage of the network? How does one maximize the throughput? This is the end-to-end Black Box problem. Prime classical approaches sampled in the sequel include TCP Tahoe, TCP Reno & TCP Vegas. Our new contribution, S-Channels, is briefly introduced and discussed. It is worthwhile to note that Black Box modeling as introduced here is in fact part of a far deeper, more fundamental problem, discussed elsewhere. (Jagannathan, 2003). White Box: This approach imposes structure on network routers. It consists of adjusting router behaviour, in terms of managing router queues and buffers to maximize network service to end users. Common examples of White Box models include the RED family of routers and Morris routers. As will be seen, Black Box models operate at the Transport Layer (Layer 4), and White Box models are at the Network Layer (Layer 3) of the TCP/IP protocol stack. In Black Box we optimize network usage, whereas in White Box we optimize the network itself. White box modeling is beyond the scope of this paper.

3 Conventional TCP Congestion Control

The disciplines in this school of thought revolve more or less around the interplay of a few key concepts or notions which are itemized below.

Congestion Windows (cwnd)
Assume the network has a capacity of delivering C bytes/sec. Assume the round trip time between source and destination is RTT. If one waited for individual packets to be acknowledged before transmitting the next one, one can send at most 1 packet per RTT. The alternative is to send cwnd = C x RTT packets spread over an RTT and wait for acks to slide this window. This allows one to send cwnd packets per RTT, and hence is more optimal. This whole argument hinges on the assumption that the capacity C is a key characteristic of the network.

TCP connections
This is the idea of an end-to-end conversation between source and destination, regardless of the "route " taken by the data constituting the conversation. TCP is at the end-to-end layer 4 (Transport layer) with the actual underlying path taken by the packets is with IP at layer 3 (Network layer). Naturally TCP connections may expect a level of Quality of Service from the underlying network layer, this interaction is beyond the scope of this paper. Conventional TCP congestion control algorithms all concern themselves with the empirical estimation of the true value of cwnd the

congestion window, and then monitoring it. The three classical congestion control algorithms are

- TCP Tahoe;
- TCP Reno;
- TCP Vegas.

We now briefly describe each. Without loss of generality, assume all packets are of fixed size = 1 byte.

3.1 TCP Tahoe

Start the window at 1. Maintain a threshold parameter (ssthresh), initially set suitably. As acknowledgements arrive from the destination (egress), double cwnd for each window-full of acknowledgements, i.e., cwnd = cwnd + 1, per ack. Do this repeatedly as long as cwnd <= ssthresh. This algorithm is called Slow Start in the literature. When cwnd > ssthresh, increase cwnd by 1 per each window-full of ACKs, i.e., cwnd = cwnd + 1/cwnd per ack. This behaviour is called Congestion Avoidance in the literature. When a timeout occurs, i.e., packet lost (due to congestion), set ssthresh = cwnd/2, and also set cwnd = 1, to re-enter SlowStart More details on Tahoe may be found in (Stevens 2001) and (Chiu et al, 1989).

3.2 TCP Reno

Packets sent by TCP are sequentially numbered, and meant to be received (assembled) in that order as well. If a specific packet does not reach, and a subsequent one does, an ACK is immediately generated which contains the sequence number of the missing packet. When three such duplicate ACKs are received, i.e., with the sequence number of a missing packet, TCP Reno does the following ssthresh = cwnd/2 cwnd = cwnd + 3 Send the problem packet, without waiting for timeout This behaviour is called Fast Retransmit in the literature. Also cf. (Allman et al, 2001) in this connection. After this when a non-duplicate Ack arrives, TCP Reno sets
cwnd =ssthresh
Then, enter Congestion Avoidance. This behaviour is called Fast Recovery in the literature.
Note that TCP Reno only works when cwnd >= 4. Also both Tahoe and Reno use ACK-clocking wherein the arrival of ACKs clocks out the new transmissions Further information on Reno can be found in (Stevens, 2001).

3.3 TCP Vegas

Fix a & b = Vegas parameters.

Compute Expected throughput as Cwnd/BaseRTT [BaseRTT is the known RoundTripTime]

Compute Actual throughput as Cwnd/RTT [RTT is the observed RoundTripTime]

$$\text{Calculate Diff} = (\text{Expected - Actual}) \times \text{BaseRTT}$$

$$Cwnd = \begin{cases} Cwnd + 1 & \text{if Diff} < \text{a} \\ Cwnd - 1 & \text{if Diff} > \text{b} \\ Cwnd, & \text{otherwise} \end{cases}$$

It may be noted that Vegas uses fluctuation in Round Trip Times to make inferences about congestion in the network. TCP Vegas was introduced in (Brakmo et al, 1995).

4 Comparison/Evaluation of Tahoe, Reno & Vegas, and Further Developments

There have been a fair number of analytical and empirical investigations that evaluate the relative efficiency, fairness and stability [steady-state behaviour] of Tahoe, Reno and Vegas. Most of these analyses use the basic model of two competing TCP connections (each running a classical congestion control protocol) sharing a wired link between two routers, or at most a simple variant of this scenario, which then lends itself to mathematical/statistical analysis. (Hasegawa et al, 1999) find that for this model Tahoe & Reno are biased towards TCP connections with longer Round Trip Times (RTTs). As a consequence of this, given two connections with the same RTT each, Tahoe & Reno are fair towards them. Clearly, by their very nature, Reno & Tahoe never stabilize, in the long run, with oscillating window sizes. Vegas, on the other hand, does stabilize in the steady-state, but the throughputs of two competing connections [with or without the same RTTs] sometimes converge differently. In words, Vegas cannot guarantee fairness. Thus, they conclude that Fairness and Stability cannot be simultaneously achieved. On the other hand, when Reno & Vegas run concurrently, (Mo et al,1999) note thatReno over-utilizes bandwidth as compared to Vegas. They posit, therefore, that despite Vegas' reported effiency being 37-71 % more than Reno (Ahn et al, 1995), this may be a reason why Vegas has not been widely deployed, i.e., it cannot compete with concurrent Reno connections. We note here that the problem with most comparative analyses of Tahoe, Reno and Vegas is that they mix

Black Box and White Box modeling. As a result any theorem that is derived in a specific network model setting is at best a counter-example, and not a general solution to the problem. Our approach, on the other hand, is Black Box model: Optimize network usage regardless of the internal structure/behaviour of the network components. Our specific model for this is S-Channels. Then,White Box model: Optimize the network itself (out of scope for this paper).
Some additional insights into S-Channels Tahoe/Reno use Timeouts and duplicate ACKs to infer about congestion, and offer algorithms to deal with that congestion. This has been discussed. On the other hand, Vegas uses fluctuations in RTT (Round Trip Time) as a measure of congestion and offers some means to counter it. Our position is that neither Tahoe/Reno (losses) of Vegas (RTT variations) is the best way to deal with congestion. In (Jagannathan, 2003) we present a different approach, S-Channels, and evaluate it empirically using WAN simulation across the Internet. In summary, some of S-Channels' key characteristics are:

- Infer about congestion based on ingress/egress rates,
- Eliminate ACKs and conserve ACK space in the TCP header,
- Manage losses efficiently,
- Be more effective than the other schemes and finally,
- Maximize throughput in the steady-state.

5 Conclusion

In this conference paper, we introduced the notion of congestion in internetworks. A scheme for managing congestion was presented, in terms of Black Box and White Box models. We surveyed leading Black Box models, and examined their pros and cons. Our new contribution, S-Channels, was outlined, along with its key features, characteristics and advantages.

References

Ahn, J.S. et al (1995). Evaluation of TCP Vegas:emulation & experiment. *IEEE Transactions in Communications*, **25**(4), 185–195.

Allman, M. et al (2001). Enhancing TCP's loss recovery using limited retransmit. *RFC*, **3042**.

Brakmo, L.S. et al (1995). TCP Vegas: End-to-end congestion avoidance in a global internet. *IEEE Journal in Selected areas in Communications*, **13**(8), 1465–1480.

Chiu, D-M. et al (1989). Analysis of the Increase & Decrease Algorithms for congestion avoidance in computer networks. *Computer Networks & ISDN systems*, **17**, 1–14.

Hasegawa, G. et al (1999). Fairness & Stability of Congestion Control mechanisms of TCP. *IEEE*.

Jagannathan, R.S. et al. A new approach to Black Box congestion management. In preparation.

Mo, J. et al (1999). Analysis and comparison of TCP Reno & Vegas. *IEEE*.

Stevens, W.R. (2001). TCP Slowstart, Congestion Avoidance, Fast Retransmit & Fast Recovery. *Internet RFC*.

Modelling Strategies for Longitudinal Data with Missingness.

Ivy Jansen[1] and Geert Molenberghs[1]

[1] Biostatistics, Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, Building D, B-3590 Diepenbeek, Belgium , ivy.jansen@luc.ac.be

Abstract: A lot of research has been devoted to modelling strategies for longitudinal data with missingness, especially within the MNAR context. In this paper, an overview will be given of several existing methods, and the relatively unexplored domain of non-monotone missingness with multivariate ordinal responses will be broached. In this context, the Multivariate Dale model (Molenberghs and Lesaffre 1994) will be used.

Keywords: Longitudinal data; Missing data; Multivariate Dale model.

1 Introduction

In applied sciences, one is often confronted with the collection of *correlated data* or otherwise hierarchical data. This generic term embraces a multitude of data structures. In particular, studies are often designed to investigate changes in a specific parameter which is measured repeatedly over time in the participating persons. Longitudinal studies are conceived for the investigation of such changes, together with the evolution of relevant covariates.
In longitudinal settings, each unit (respondent, cluster, patient,...) typically has a *vector* $\boldsymbol{Y}$ of responses. This leads to several, generally non-equivalent, extensions of univariate models. In a *marginal model*, marginal distributions are used to describe the outcome vector $\boldsymbol{Y}$, given a set $\boldsymbol{X}$ of predictor variables. The correlation among the components of $\boldsymbol{Y}$ can then be captured either by adopting a fully parametric approach or by means of working assumptions, such as in the semiparametric approach of Liang and Zeger (1986). Alternatively, in a *random-effects model*, the predictor variables $\boldsymbol{X}$ are supplemented with a vector $\boldsymbol{\theta}$ of random effects, conditional upon which the components of $\boldsymbol{Y}$ are usually assumed to be independent. This does not preclude that more elaborate models are possible if residual dependence is detected (Longford 1993). Finally, a *conditional model* describes the distribution of the components of $\boldsymbol{Y}$, conditional on $\boldsymbol{X}$ but also conditional on (a subset of) the other components of $\boldsymbol{Y}$. Well-known members of this class of models are log-linear models (Gilula and Haberman, 1994).

2 Current Practice

The analysis of longitudinal clinical trials is almost invariably hampered by dropout. In current practice methods such as *last observation carried forward* (LOCF) or *complete case* analysis (CC) are very prominent. Such less than optimal methods fall within the *missing completely at random* category (MCAR), where dropout is independent of the measurement process, and part of the literature, supported by the biopharmaceutical industry and the regulatory authorities (FDA in the United States, EMEA in Europe, and their Japanese and other national counterparts), maintains that these methods are to be preferred for reasons of simplicity and validity.
The academic research community, on the other hand, focuses to a large extent on methods for *missing not at random* (MNAR) where dropout is allowed to depend on unobserved measurements. Some researchers believe that ever more complicated MNAR methods will eventually be sufficiently general to encompass the true data generating mechanism.

3 Overview of MNAR Models for Categorical Data

In the MNAR setting, we will make a distinction between models for monotone and non-monotone missingness. The model proposed by Molenberghs, Kenward and Lessafre (1997), which combines a Dale model for the measurements with a logistic regression for dropout (as in the Diggle and Kenward (1994) philosophy), can handle monotone ordinal data. For non-monotone patterns, Baker, Rosenberger, and DerSimonian (1992) proposed a model for bivariate binary data subject to non-random nonresponse, which is reformulated by Jansen et al. (2003) using 2 loglinear models, such that its membership of the selection model family is unambiguously clear, to accommodate for, possibly continuous, covariates, turning the model into a regression tool for several categorical outcomes, and to avoid the risk of invalid solutions. A disadvantage of those BRD models, is that the parameters cannot be interpreted marginally, which is actually what clinicians want.
As we can see, until now there does not exist a model that allows for non-monotone missingness with more than 2 possible outcomes. A solution will be presented in the next section.

4 A Method for Non-monotone Categorical Outcomes

Since the multivariate Dale model (Molenberghs and Lesaffre 1999), which extends the bivariate global cross-ratio model described by Dale (1986), accounts for the dependence between multiple ordinal responses, as well

as their dependence on covariate vector(s), which may be time-varying, continuous and/or discrete, this model is very useful for our purpose.
The model arises from a decomposition of the joint probabilities into main effects (described by marginal probabilities) and interactions (described by cross-ratios of second and higher orders).
This model will be used for the measurements $\boldsymbol{Y}$ and for the missingness given the measurements $\boldsymbol{R}|\boldsymbol{Y}$, such that again a selection model is obtained and both discrete and continuous covariates can be included in the measurement model as well as in the missingness model.
Results will be presented for simulated data, and for a data set from a multicenter, postmarketing study involving 315 patients that were treated by fluvoxamine for psychiatric symptoms described as possibly resulting from a dysregulation of serotonine in the brain.

5 Need for a Sensitivity Analysis

The route of a sensitivity analysis has been explored many times in the context of categorical data. For the model by Baker, Rosenberger and DerSimonian (1992), Molenberghs, Kenward and Goetghebeur (2001) developed the intervals of ignorance and uncertainty. To the reformulated model by Jansen et al. (2003) a local influence is applied by the same authors. This local influence is also applied to the Dale model with dropout (Molenberghs, Kenward and Lessafre, 1997) by Van Steen et al. (2001). Future work will be devoted to a sensitivity analysis on the model for non-monotone categorical outcomes, that was introduced in the previous section.

References

Baker, S.G. (1995). Marginal regression for repeated binary data with outcomes subject to non-ignorable non-response. *Biometrics*, **51**, 1042–1052.

Baker, S.G., Rosenberger, W.F., and DerSimonian, R. (1992). Closed-form estimates for missing counts in two-way contingency tables. *Statistics in Medicine*, **11**, 643–657.

Dale, J.R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics*, **42**, 909–917.

Diggle, P.D. and Kenward, M.G. (1994). Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics*, **43**, 49–93.

Gilula, Z. and Haberman, S. (1994). Conditional log-liear models for analyzing categorical panel data. *Journal of the American Statistical Association*, **89**, 645–656.

Jansen, I., Molenberghs, G., Aerts, M., Thijs, H., and Van Steen, K. (2003). A local influence approach applied to binary data from a psychiatric study. *Biometrics*, **59**, 409–418.

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

Longford, N.T. (1993). *Random Coefficient Models*. London: Oxford University Press.

Molenberghs, G., Kenward, M.G., and Goetghebeur, E. (2001). Sensitivity analysis for incomplete contingency tables: the Slovenian plebiscite case. *Applied Statistics*, **50**, 15–29.

Molenberghs, G., Kenward, M.G., and Lesaffre, E. (1997). The analysis of longitudinal ordinal data with nonrandom dropout. *Biometrika*, **84**, 33–44.

Molenberghs, G., and Lesaffre, E. (1999). Marginal modelling of multivariate categorical data. *Statistics in Medicine*, **18**, 2237–2255.

Van Steen, K., Molenberghs, G., Verbeke, G., and Thijs, H. (2001). A local influence approach to sensitivity analysis of incomplete longitudinal ordinal data. *Statistical Modelling:* , **1**, 125–142.

Model Based Clustering for Multivariate Count Data

Dimitris Karlis[1] and Loukia Meligkotsidou[2]

[1] Department of Statistics, Athens University of Economics and Business, 76 Patission str, 10434, Athens, Greece and
[2] Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, United Kingdom

Abstract: Model based clustering has found increasing applications in recent years in contrast to ad-hoc and not statistically solid methods, like hierarchical clustering and K-means method. Model based clustering has been created on probabilistic grounds that assume the existence of a finite mixture model and transform the problem to one of estimating the parameters of the multivariate mixture model. Several inferential procedures have been described for this model, including model selection and prediction among others. In the present paper we propose model based clustering for count data based on the multivariate Poisson distribution which is the natural counterpart of the multivariate normal model. The general formulation of the model as well as estimation procedures are provided. Finally, problems for further research are discussed.

Keywords: Multivariate Poisson distribution; EM algorithm; Crime data; Finite mixtures

1 Introduction

Multivariate count data appear in a wide range of fields like epidemiology (e.g. different type of a disease), marketing (e.g. purchases of different products), environmentrics (e.g. different kind of plantation etc) just to name few. While the Poisson distribution has played a prominent role in modelling univariate data, its multivariate counterpart has been rarely used in practice mainly due to computational difficulties for inferential procedures (see, e.g. Johnson et al, 1997). The use of multivariate normal models as approximation to the multivariate Poisson model, can be misleading especially if the means are small and there are a lot of zero counts.
The multivariate Poisson distribution, while the most important among discrete multivariate distributions (see, e.g. Johnson et al, 1997), has several shortcomings for its application. The main drawback of the application of the multivariate Poisson distribution is the complicated form of the joint probability function that has led to the use of a simplified model with just a common covariance term for all the pairs of variables (see Tsionas, 1999, 2001 and Karlis, 2003).

In the present paper a multivariate Poisson distribution with different covariances for all the pairs of variables will be described. A finite mixture model of this family of multivariate Poisson distributions will be constructed aiming at clustering multivariate count observations

2 Model Based Clustering

Historically, cluster analysis has developed mainly through ad-hoc methods based on empirical arguments. The last decade, however, there is an increased interest in model based methodologies, which allow for clustering procedures based on statistical arguments and methodologies. The majority of such procedures are based on the multivariate normal distribution (see, for instance, Banfield and Raftery, 1993; McLachlan and Basford, 1988). The central idea of model-based clustering is the use of finite mixtures of a density. The population of interest, thus, consists of k subpopulations and the density (or probability function) of the multidimensional observation $\boldsymbol{y}$ from the j-th subpopulation is $f(\boldsymbol{y}|\theta_j)$ for some unknown vector of parameters θ_j. Since, we do not observe the cluster labels, the unconditional density of the vector y is a mixture density of the form

$$f(\boldsymbol{y}) = \sum_{j=1}^{k} p_j f(\boldsymbol{y} \mid \theta_j) \tag{1}$$

where $0 < p_j < 1, \sum p_j = 1$ are the mixing proportions. Note that the mixing proportion is the probability that a randomly selected observation belongs to the j-th cluster. This is the classical mixture model (see, e.g. Bohning, 1999; McLachlan and Peel, 2000). The purpose of model based clustering is to estimate the parameters $(\theta_1, \ldots, \theta_k, p_1, \ldots, p_{k-1})$. An expectation-maximization (EM) algorithm is applicable for finding ML estimates. The majority of model based clustering is based on the multivariate normal distribution and hence it is based on the assumption of continuous data. We will now propose a model adequate for multivariate count data.

3 A Multivariate Poisson Distribution

The general multivariate Poisson distribution is based on the following multivariate reduction scheme. Assuming Y_r, $r = 1, ..., k$, are independent univariate Poisson random variables, i.e. $Y_r \sim Po(\theta_r)$, $r = 1, ..., k$, then the definition of multivariate Poisson models is made through the vector $\mathbf{Y}' = (Y_1, Y_2, ..., Y_k)$ and an $m \times k$ matrix $\mathbf{A}$ with zeroes and ones. Specifically, the vector $\mathbf{X}' = (X_1, X_2, ..., X_m)$ defined as $\mathbf{X} = \mathbf{AY}$ follows a multivariate Poisson distribution.

In general, matrix $\mathbf{A}$ has the form $\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots \boldsymbol{A}_m]$, where $\mathbf{A}_j$, $j = 1, \ldots, m$ is a sub-matrix of dimensions $m \times \begin{pmatrix} m \\ j \end{pmatrix}$, each column of $\mathbf{A}_j$ has exactly j ones and $(m - j)$ zeroes and no duplicate columns exist. Thus, $\mathbf{A}_m$ is the column vector of $\mathbf{1}$s, while $\mathbf{A_1}$ becomes the identity matrix of size $m \times m$. We can realize that, since the random variable Y_k appears in every element of $\boldsymbol{X}$ can be interpreted as an m-way covariance effect to each of the X_is.

The reduced models for m variables derived from $\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_m]$ are frequently used in the literature and the resulting distributions are commonly referred to as the multivariate Poisson distributions (see, e.g. Tsionas, 1999, Karlis, 2003). This model assumes that all the covariances are the same, which is not at all realistic. We focus on the case where only the main effects and the two-way covariance effects are considered, i.e. $\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2]$, for the analysis of multivariate data sets. This is done in order not to impose too much structure to our data.

It holds that

$$E(X) = \mathbf{AM}$$

and

$$Var(X) = \mathbf{A\Sigma A}'$$

where $\mathbf{M}$ and $\mathbf{\Sigma}$ are the mean vector and the variance covariance Σ is diagonal because of the independence of Y_i's and has the form

$$\mathbf{\Sigma} = diag(\theta_1, \theta_2, \ldots, \theta_m)$$

Similarly

$$\mathbf{M} = (\theta_1, \theta_2, \ldots, \theta_m)'$$

Another interesting feature of this model is that it allows for covariance terms separately for each pair of variables and thus it can be considered as a counterpart of the multivariate normal distributions suitable for multivariate count data.

In the case of the Multivariate Poisson Distribution, the calculation of the probability mass function can be of great difficulty, as it often demands summations over high-dimensional spaces. If we consider the multivariate Poisson model with complete specification, the probability function for the m-variate case needs $\sum_{j=1}^{m} \begin{pmatrix} m \\ j \end{pmatrix} - m = 2^m - m - 1$ summations.

Fortunately, computation of the probabilities can be accomplished via recursive schemes. Kano and Kawamura (1991) provided a general scheme for constructing recurrence relations for multivariate Poisson distributions. Even those recursive relations must be used efficiently in order to lead to feasible calculations.

Note that, in the case of $\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2]$, each row of $\boldsymbol{A}$ contains exactly m ones, so each recurrence relationship for the calculation of $P(\boldsymbol{X} = \boldsymbol{x})$

requires the computation of m previous probabilities. Obviously, as m increases the complexity of the model, and hence the computational effort, increases too.

4 Finite Mixture Model

Naturally we may define a multivariate Poisson mixture model by considering a multivariate Poisson mixture where $f(\boldsymbol{y})$ in (1) is replaced by the joint probability function of the multivariate Poisson distribution.
For estimation purposes an EM type algorithm can be used. It is interesting to note that this scheme is similar to the standard approach for mixture models described in McLachlan and Peel (2000) but now one must make also use of the multivariate reduction technique upon which the derivation of the distribution is based. Thus the latent variables are the vectors Z_i where $Z_{ij} = 1$ or 0 if the i-th observation belongs to the j-th cluster or not, accordingly, together with the unobserved variables Y_{ij}. An EM algorithm for the simple case of the multivariate Poisson distribution with the covariance structure as described in the present paper is given in Karlis and Meligotsidou (2003). Inference on the model can be made via standard techniques for finite mixture models. Criteria for assessing the number of clusters can be also used.

5 Application

The real data application concerns crime data taken from National Statistical Service of Greece for the year 1996. Four different kinds of crimes were examined for 50 prefectures of Greece. The different crimes considered were rapes, arson , smuggling of antiquities and general smuggling. The population of each prefecture is used as an offset. The aim is to cluster the prefectures according to their profiles in those types of crimes.
An EM type algorithm was used to fit the finite multivariate Poisson mixture model. The number k of components was considered as known for using the EM algorithm, but we fitted the model with increasing value of k in order to decide about the number of components. For a model with k components there are $11k - 1$parameters to estimate. The AIC criterion selected the solution with 4 components and mixing proportions $\hat{p} = (0.5915, 0.2266, 0.0638, 0.1181)$. The parameters for each component are given below in a matrix form, where the diagonal parameters are the mean parameters while the non-diagonal elements are the covariances between the pairs. Note that the marginal means are in fact the sum of the so-called mean parameter and the covariance parameters related to the variable.

$$\Theta_1 = \begin{bmatrix} 17.339 & 0 & 4.772 & 0 \\ & 2.530 & 0.198 & 1.977 \\ & & 34.112 & 2.398 \\ & & & 6.171 \end{bmatrix}, \Theta_2 = \begin{bmatrix} 0 & 0 & 3.675 & 0 \\ & 9.897 & 1.925 & 1.938 \\ & & 5.320 & 0 \\ & & & 2.172 \end{bmatrix},$$

$$\Theta_3 = \begin{bmatrix} 0 & 20.424 & 0 & 0 \\ & 55.868 & 24.323 & 0 \\ & & 0 & 0 \\ & & & 0 \end{bmatrix}, \Theta_3 = \begin{bmatrix} 14.416 & 12.780 & 0 & 0 \\ & 3.048 & 0 & 0 \\ & & 8.934 & 0 \\ & & & 44.621 \end{bmatrix}$$

It is interesting to see that the structure of the parameters is quite different for the four components. The third component is associated with high rate of smuggling of antiquities, in fact it consists of a single prefecture. The 4th component has not any covariances apart from the covariance between rapes and arsons. The first two components, which are the largest ones, have enough structure with large covariances between variables. Thus our approach decompose in some way the entire covariance between the variables to different parts. More details will be given in a forthcoming complete version of the paper.

References

Bohning, D. (1999). *Computer Assisted Analysis of Mixtures and Applications*. New York: Chapman & Hall.

Banfield, J.D. and A.E. Raftery (1993). Model based gaussian and non-gaussian clustering. *Biometrics*, **49**, 803–821.

Johnson, N., Kotz, S., and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. New York: Wiley.

Karlis, D. (2003). An EM algorithm for multivariate poisson distribution and related models. *Applied Statistics*, **30**, 63–77.

Karlis, D. and Meligkotsidou, L. (2003.) Multivariate poisson regression with full covariance structure. Submitted.

Kano K. and Kawamura, K. (1991). On recurrence relations for the probability function of multivariate generalized poisson distribution. *Communications in Statistics -Theory and Methods*, **20**, 165–178.

McLachlan, G. and K. Basford (1988), *Mixture Models: Inference and Applications to Clustering*. New-York: Dekker.

McLachlan, G. and D. Peel (2000), *Finite mixture models*. New York: Wiley.

Tsionas, E.G. (1999). Bayesian analysis of the multivariate poisson distribution. *Communications in Statistics - Theory and Methods*, **28**, 431–451.

Tsionas, E.G. (2001). Bayesian multivariate poisson regression. *Communications in Statistics - Theory and Methods*, **30**, 243–255.

Penalised Spline Smoothing in Multivariable Survival Models with Varying Coefficients

Göran Kauermann[1] and Denise Brown[2]

[1] Universität Bielefeld, Fakultät für Wirtschaftswissenschaften, Postfach 10 01 31, 33501 Bielefeld, Germany, gkauermann@wiwi.uni-bielefeld.de
[2] University of Glasgow, Department of Statistics, Glasgow G11 8QQ, Scotland, denise@stats.gla.ac.uk

Keywords: Mixed models; Penalised smoothing, Survival models, Varying coefficient model.

1 Introduction

Modelling of survival data is largely dominated by the proportional hazard model introduced by Cox (1972). Even though the model is appealing, the proportional hazard (PH) assumption is often not fulfilled in applications when covariate effects vary with survival time. The assumption has been under major investigation and numerous papers suggest test procedures or extensions, see for instance O'Sullivan (1988), Hastie and Tibshirani (1990), Gray (1994), Hess (1994) or Abrahamowicz et al (1996). Allowing covariate effects to vary with time leads to a varying coefficient model as generally introduce by Hastie and Tibshirani (1993). Here, constant covariate effects are replaced by smooth but unknown function. Smooth estimation can then be carried out using e.g. Spline fitting, as in Hastie and Tibshirani (1993), or applying local techniques, as e.g. in Fan et al (1997).
In this paper we employ penalised spline fitting (P-spline) as smooth estimation procedure. The approach has been originally introduced by O'Sullivan (1986), but the procedure finally achieved recognition due to the paper by Eilers and Marx (1996). The approach is numerically very handy and uncovers strong similarities to penalised quasi likelihood estimation in Generalised Linear Mixed Models, as discussed in Breslow and Clayton (1993). This link becomes obvious if the penalty is rewritten as prior distribution on the coefficients of the basis. In fact, the smoothing parameters steering the amount of penalty plays the role of the variance in the Generalised Linear Mixed Model formulation. We demonstrate how this connection can be used to estimate the smoothing parameter appropriately. A general discussion about the connection of P-spline smoothing and Mixed Models is also found in Wand (2003).

The P-spline approach pursued in this paper is imposed directly on the likelihood, rather than on the partial likelihood. This not only allows to smoothly estimate the baseline hazard, it also follows technically more closely the likelihood principle in the case of non-proportional hazards. In particular the integrated hazard function in the likelihood is approximated using a trapezoid integration. This in turn leads to simple likelihood functions resembling a Poisson model.

2 Smooth Hazard Model

2.1 *P*-Spline Fitting

Let T_i denote the survival time of the ith individual or observational unit and let C_i be the corresponding right censored time, $i = 1, \ldots, N$. We observe $Y_i = \min(T_i, C_i)$ and define the censoring indicator $\delta_i = 1$ if $T_i < C_i$ and $\delta_i = 0$ otherwise. With x_i we denote the p dimensional covariate vector for the i-th individual, which for simplicity of notation is assumed to be time constant. The hazard function is then modelled as

$$h(t, x_i) \quad = \quad \lambda_0(t) \exp\{x_i^T \beta_x(t)\} \tag{1}$$

with $\lambda_0(t)$ as baseline hazard and $\beta_x(t)$ as vector of covariate effects varying smoothly with survival time t. For convenience we rewrite (1) to $h(t, x_i) = \exp\{z_i \beta(t)\}$ with $z_i^T = (1, x_i^T)$ and $\beta(t) = \{\log \lambda_0(t), \beta_x^T(t)\}^T$. The task is to estimate $\beta(t)$ smoothly by avoiding any stringent parametric assumptions. This is achieved by penalised spline regression.

For the sake of simplicity let us first consider smooth estimation of the baseline function $\beta_0(t) = \log \lambda_0(t)$. Let $B(t) = \{b_1(t), \ldots, b_q(t)\}$ be a basis developed over the knots $t_1, \ldots, t_q$. A convenient choice is to use a B-spline basis (see Boor, 1978), even though other choices are possible as well. The dimension q of the basis is chosen lavish, such that the model bias $\beta_0(t) - B(t)\alpha_0^0$ is negligible, where $\alpha_0^0 = (\alpha_{01}^0, \ldots \alpha_{0q}^0)^T$ is the vector of "best" coefficients in the sense of having minimal Kullback-Leibler distance. Since q is supposed to be large, simple maximum likelihood estimation of α_0 would be highly variable and numerically unstable. Therefore, in order to achieve smoothness and numerical stability the penalty term $\lambda_0 \alpha_0^T D_0 \alpha_0$ is introduced, with D_0 as appropriately chosen penalty matrix and λ_0 as bandwidth steering the amount of penalisation.

In the same fashion we fit the remaining components in the model. It is thereby tactically an advantage to extract the intercept from the smooth function. This means for estimation we decompose $\beta_l(t)$ to $\beta_{0l} + \tilde{B}(t)\alpha_l$, $l = 0, \ldots, p$, with $\alpha_l = (\alpha_{l1}, \ldots, \alpha_{lq})^T$ and $\tilde{B}(t)$ as basis matrix containing no intercept. We define $\theta_l = (\beta_{0l}, \alpha_l^T)^T$ and using the Kronecker product we can jointly write

$$\beta(t) \quad = \quad \boldsymbol{W}(t)\boldsymbol{\theta}$$

with $\boldsymbol{W(t)} = I_{p+1} \otimes \{1, \tilde{B}(t)\}$ and parameter vector $\boldsymbol{\theta}^T = (\theta_0^T, \ldots, \theta_q^T)$, where I_{p+1} is the $p+1$ dimensional identity matrix.
Coefficients α_l are now jointly penalised to achieve smooth fits. This leads to the penalised likelihood

$$l^P(\boldsymbol{\theta}, \lambda) \quad = \quad \sum_{i=1}^{N} l_i(\boldsymbol{\theta}) - \sum_{l=0}^{p} \lambda_l \alpha_l^T D_l \alpha_l \tag{2}$$

with $l_i(\boldsymbol{\theta}) = \delta_i \Big(z_i^T \boldsymbol{W}(Y_i)\boldsymbol{\theta} \Big) - \int_0^{Y_i} \exp\{z_i^T \boldsymbol{W}(t)\boldsymbol{\theta}\} \mathrm{d}t$ as likelihood contribution and $\lambda = (\lambda_0, \ldots, \lambda_p)$ as component-wise smoothing parameters steering the amount of penalisation for each component. For notational convenience the penalty component can be rewritten to $\boldsymbol{\theta}^T(\Lambda \boldsymbol{D})\boldsymbol{\theta}$ with $\boldsymbol{D}$ as block diagonal matrix build from matrices D_l and zero entries for β_{0l}, $l = 1, \ldots, p$. Bandwidth matrix $\boldsymbol{\Lambda}$ matches accordingly as diagonal built from λ_l
Differentiating (2)with respect to $\boldsymbol{\theta}$ leads to the penalised score equation

$$\frac{\partial l^p(\boldsymbol{\theta}, \lambda)}{\partial \boldsymbol{\theta}} \quad = \quad \sum_{i=1}^{N} \boldsymbol{s}_i(\boldsymbol{\theta}) - \Lambda \boldsymbol{D} \boldsymbol{\theta} \quad = \quad 0 \tag{3}$$

with $\boldsymbol{s}_i(\boldsymbol{\theta}) = \delta_i \boldsymbol{W}^T(Y_i) z_i - \int_0^{Y_i} \boldsymbol{W}^T(t) z_i \exp\{z_i^T \boldsymbol{W}(t)\boldsymbol{\theta}\} dt$. Accordingly, the second order derivative results to

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \quad = \quad \sum_{i=1}^{N} \nabla \boldsymbol{s}_i(\boldsymbol{\theta}) - \Lambda \boldsymbol{D} \tag{4}$$

where $\nabla \boldsymbol{s}_i(\boldsymbol{\theta}) = -\int_0^{Y_i} \boldsymbol{W}^T(t) z z_i^T \boldsymbol{W}(t) \exp\{z_i^T \boldsymbol{W}(t)\boldsymbol{\theta}\} dt$.

2.2 Integration

The penalised likelihood and its derivatives (3) and (4) contain integrals based on the hazard function and its derivatives. Since no analytic solution for the integrals exist, numerical integration is required. A computationally handy version is to approximate the integrals by trapezoids. It can be shown that this leads to likelihood contributions resulting from Pseudo Poisson variables $\tilde{Y}_{ij}$, say. In particular this means that standard software can be used for fitting.

3 Relation to Generalised Linear Mixed Models

Penalised spline smoothing has strong affinities to penalised quasi likelihood estimation in Generalised Linear Mixed Models (GLMM) as discussed in

Breslow and Clayton (1993). For normal response models this link is illuminated in depth in Wand (2003). For non-normal response models such link is achieved in close analogy. In particular we consider coefficients α_l, $l = 0, \ldots, p$, as independent normally distributed variables with

$$\alpha_l \sim N(0, \lambda_l^{-1} D_l^-) \tag{5}$$

where D_l^- is a generalised inverse of D_l. The bandwidth parameters λ_l now occurs in the *a priori* variance of α_l. Conditional on α_l, $l = 0, \ldots, p$ and based on the trapezoid integration we model the Pseudo Poisson variables as

$$\tilde{Y}_{ik} | (\alpha_0, \ldots, \alpha_p) \sim Po(z_i^T \boldsymbol{B}(\tau_k)\boldsymbol{\theta} + o_{ik}) \tag{6}$$

with o_{ik} as know offset resulting from the trapezoid integration. Apparently, (5) and (6) provide the ingredients of a Generalised Linear Mixed Model. The likelihood for parameters β_{0l} and λ_l, $l = 0, \ldots, p$, is obtained by integrating out the random coefficients, i.e.

$$\begin{aligned} l(\beta_{00}, \ldots, \beta_{0p}, \lambda_0, \ldots, \lambda_p) &= \int \prod_{i=1}^{N} \prod_{k=1}^{K_i} Po(\tilde{Y}_{ik}; z_i^T \boldsymbol{B}(\tau_k)\boldsymbol{\theta} + o_{ik}) \\ &\quad \times \prod_{l=0}^{p} \phi(\alpha_l, \lambda_l^{-1} D_l^-) \mathrm{d}\alpha_l \end{aligned} \tag{7}$$

with $\phi(\cdot)$ as normal density. Using Laplace approximation for the integral leads to penalised quasi likelihood estimation. It is not difficult to derive that this in turn is equivalent to the penalised estimating equations for the original survival model.

The connection between smoothing and GLMMs is not only of theoretical nature but can be exploited practically to choose an appropriate bandwidths λ_l, $l = 0, \ldots, p$. The idea is to estimate λ_l based on the likelihood function (7). Approximating the integral by Laplace integration and inserting estimates for β_{0l} provides a Laplace approximation for the log profile likelihood given by

$$l^P(\lambda_0, \ldots, \lambda_p) = \sum_{i=1}^{N} \sum_{k=1}^{K_i} \log Po(\tilde{Y}_{ik}; \cdot) - \frac{1}{2} \sum_{l=0}^{p} \left(\hat{\alpha}_l^T D_l \hat{\alpha}_l + \log |\lambda_l D_l| \right).$$

Maximising this with respect to λ_l gives

$$\hat{\lambda}_l = \frac{q}{\hat{\alpha}_l^T D_l \hat{\alpha}_l}. \tag{8}$$

as corresponding estimate. Apparently, $\hat{\lambda}_l$ depends on $\hat{\alpha}_l$ and vice versa. Therefore a simple backfitting type argument has to be applied. Cycling between (3) and (8) gives the final estimate. The estimates for λ_l can now be employed for model selection, that is if $\hat{\lambda}_l$ is large (i.e. $\hat{\lambda}_l \to \infty$) there is no evidence that the l covariate has a time varying effect.

4 Application

In the talk we demonstrate the approach by an application. Moreover simulations are provided to illuminate the model selection aspect.

References

Abrahamowicz, M., MacKenzie, T., and Esdaile, J.M. (1996). Time- dependent hazard ratio: Modeling and hypothesis testing with application in lupus nephritis. *Journal of the American Statistical Association,* **91**, 1432–1439.

de Boor, C. (1978). *A Practical Guide to Splines*. Berlin: Springer.

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

Cox, D.R. (1972). Regression models and life tables (with discussion) *Journal of the Royal Statistical Society, Series B,* **34**, 187–220.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with *P*-splines and penalties. *Statistical Science,* **11**(2), 89–121.

Gray, R.J. (1994). Spline-based tests in survival analysis. *Biometrics*,**50**, 640–652.

Hastie, T. and Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazard model. *Biometrics*, **46**, 1005–1016.

Hastie, T. and Tibshirani, R. (1993). Varying–coefficient models. *Journal of the Royal Statistical Society, Series B,* **55**, 757–796.

Hess, K. R. (1994). Goodness-of-fit tests for the general Cox regression model. *Statistica Sinica,* **1**, 1–17.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (c/r: P519-527). *Statistical Science*, **1**, 502–518.

O'Sullivan, F. (1988). Nonparametric estimation of relative risk using splines and cross-validation. *SIAM Journal on Scientific and Statistical Computing,* **9**, 531–542.

Wand, M. (2003). Smoothing and mixed models. *Computational Statistics.* (in press).

Calibration of NIR Spectroscopy Instruments: A Comparison of Various Statistical/AI Modelling Techniques

Martin Kidd[1]

[1] Centre for Statistical Consultation, University of Stellenbosch, Private Bax X1, Matieland 7602, South Africa. Email: mkidd@sun.ac.za

Abstract: Near infrared (NIR) spectroscopy instruments are used as a non-destructive method for determining and predicting various characteristics of foodstuffs. Statistical methods are used to calibrate the spectroscopy instruments. Partial Least Squares Regression is the preferred technique currently used. This paper describes a comparative study where various statistical and artificial intelligence techniques were compared with the more traditional methods in terms of prediction power. The results indicated that multivariate adaptive regression splines and support vector machines are superior to partial least squares and ridge regression.

Keywords: PLS regression; MARS; Support vector machines; Ridge regression; NIR spectroscopy.

1 Introduction

Near infrared (NIR) spectroscopy instruments are used as a non-destructive method for determining and predicting various characteristics of foodstuffs. The foodstuffs are illuminated with a range of frequencies in the near-infrared region. Based on the amount of light energy returned, the absorption of energy can be determined for each frequency used. The characteristics of the foodstuffs (eg water, sugar, fat contents) absorb different amounts of light energy at different frequencies. Therefore, by analysing the returned energy at the various frequencies, the characteristics of the foodstuffs can be predicted without having to analyse the food in a laboratory.

In order for the instrument to be able to predict the characteristics of the food, it first has to be calibrated. What is meant by calibration is that a statistical model is fitted on a set of spectroscopy data, and then subsequently used for prediction purposes. The data set consists of a set of NIR frequencies (predictor variables) and one or more values quantifying the characteristics of the food under study (target vari-

ables). The values for the target variables are determined in the laboratory.

The calibration is a three step process of pre-processing of the raw data, fitting the statistical model and evaluating the effectiveness of the fitted model. For this purpose the data is split into a training set and a test set. The model parameters are estimated from the training set and the effectiveness determined from the test set. Measures of effectiveness include the MSE (mean square error) or the correlation between estimated and actual values of the target variables.

The problem that arises with the data is that of multicollinearity. The independent variables are highly correlated and this renders the application of ordinary least squares regression impossible. Various other techniques exist which can be used to circumvent this problem. In the chemometrics field the method of partial least squares (PLS) has become the standard for calibrating NIR instruments.

Comparative studies have been done in the past to compare various techniques with one another in the role of calibrating NIR instruments, which included PLS, Principal Component Regression (PCR), and Ridge Regression. In the last 10 years other techniques for building statistical models have come to the fore, either developed by statisticians or artificial intelligence researchers. These techniques include Neural Networks, Regression Trees, Multivariate Adaptive Regression Splines (MARS) and Support Vector Machines (SVM). No reference could be found in the literature where these techniques were evaluated in a NIR calibration role.

A comparative study was done to compare the above-mentioned techniques. In section 2 the method for comparison of the techniques is described. In section 3 the data sets used for comparison are discussed. Section 4 briefly describes the techniques included and the results are summarised in section 5.

2 Method for Comparison

The techniques were compared using two actual data sets from the field of chemometrics. Each of the data sets had one target variable to be estimated using NIR calibration. These data sets are discussed in more detail in section 3.

The data was divided into a training set and a test set by randomly selecting 80% of the data (without replacement) for the training set and the remaining for the test set. Calibration models were derived for each of the techniques from the training set and applied to the test set. The mean

square error (MSE) was calculated from the test set results for comparing prediction accuracy. The MSE was calculated as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where n is the number of cases in the test set, y_i the actual value of the dependent variable and $\hat{y}_i$ the estimate from the calibration model.

The above process was then repeated 10 times resulting in 10 different MSE values for each technique. Box plots were constructed for the MSE values for comparative purposes.

3 Data Sets

The first data set originated from a study where NIR calibration was applied to meat samples for the purpose of estimating the fat content of the meat. A total of 94 samples were scanned and Figure 1 shows a typical graph of absorption versus frequency for one of the meat samples. An empirical first derivative was calculated at each frequency and included as predictor variables.The fat content for each of the 94 samples were determined in a laboratory, which served as the target variable. This data set was used courtesy of Claus Borgaard from the Danish Meat Institute.

The second data set was generated by an experiment to predict the soluble solid content (brix values) of peaches. Figure 1 shows a typical graph of frequency vs absorption for one of the peaches scanned. As with the first data set, empirical first derivatives were also calculated for this data set and added to the set of predictor variables. This data set was used courtesy of the Post-Harvest & Wine Technology Division, ARC Infruitec-Nietvoorbij, and the Dept of Food Science at Stellenbosch University, South Africa.

4 Modelling techniques included in the study

The following techniques were included in the comparative study:

- Partial Least Squares Regression (PLS)
- Ridge Regression
- Multivariate Adaptive Regression Splines (MARS)
- Support Vector Machines (SVM)

The basic principles of each of the techniques are discussed in the following paragraphs.

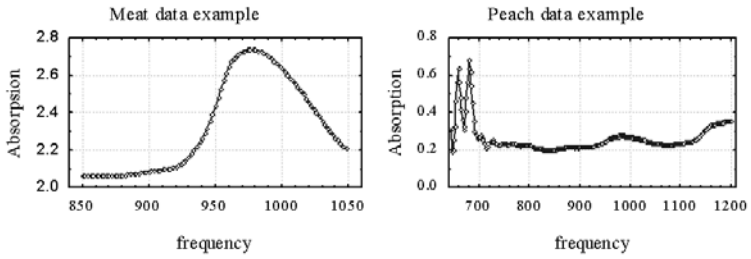


FIGURE 1. *Plot of absorption vs frequency for one of the meat and peach samples.*

4.1 Partial Least Squares

Partial least squares regression (PLS) is similar to principal components regression(PCR). In PCR the dimension of the input matrix (independent variables) is reduced by extracting principal components. Regression is then performed by using the principal component scores as inputs. PLS extends this idea by using the target variable in addition to the input variables for constructing the principal component scores. All PLS calculations were done using Statistica v6.

4.2 Ridge Regression

Ridge regression is specifically used when the independent variables are highly correlated and stable estimates for the regression parameters cannot be obtained. It artificially decreases the correlations so that more stable but biased estimates of the regression coefficients can be computed. This is achieved by adding a constant (α) to the diagonal elements of the correlation matrix and then re-standardising the diagonal elements to one (the off-diagonal elements are divided by the constant). All ridge regression calculations were done using Statistica v6.

4.3 Multivariate Adaptive Regression Splines

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. Each of the regression lines operate on distinct non-overlapping regions of the predictor variable space. The position where one regression line stops and the next line starts, is called a knot position. MARS derives the knot positions from the data. It can also handle more than one predictor variable as well as combinations of categorical and continuous predictors. From a MARS analysis it is possible to determine the relative importance of predictor variables with respect to the target

variable. All MARS calculations were done using MARS v2 from Salford Systems.

4.4 Support Vector Machines

Support vector machines(SVM) are better known for application to classification problems. In the classification setting the 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. This method is then adapted for the regression case, keeping some properties of the SVM classifier. All SVM calculations were done using R and functions written by David Meyer (based on C/C++-code by Chih-Chung Chang and Chih-Jen Lin).

5 Results

All the techniques included have tuning parameters which can be varied to find the best fit. Thus, before the techniques were compared with one another, tuning parameters giving the best fit for each of the data sets were derived.

Figure 2 gives the results for the meat data set. From the graph we see that MARS produced the best results with SVM also performing better than the traditional methods of ridge regression and PLS. A plot of the residuals for PLS showed indications of a non-linear relationship between the target and predictor values. The above results show the ability of MARS and SVM to model this non-linearity.

Figure 2 also shows the comparative performances for the peach data. From the graph we see that there is not much difference between the techniques (contrary to the meat data), but MARS still appeared to provide better results than the other techniques. Inspection of the PLS residuals did not indicate non-linear relationships between the target and predictor variables.

6 Conclusion

The results from Figure 2 indicate that MARS and SVM are better techniques to use for calibration of NIR instruments than PLS or ridge regression. In cases where non-linearity was not present in the data, the performance of MARS and SVM were similar to PLS (with MARS showing a slight improvement), but where there were non-linear relationships in

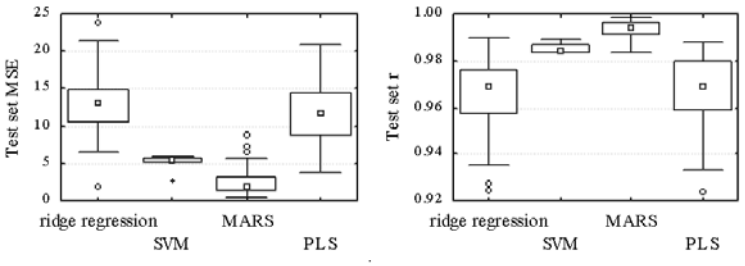


FIGURE 2. *Comparison of the techniques for the meat and peach data.*

the data, MARS and SVM clearly outperformed PLS and ridge regression.

A further advantage of MARS is that it that the software provides relative importance scores for the predictor variables. Thus MARS have the ability to highlight important frequencies having the most predictive power.

This study will be extended by including two more techniques namely neural networks and projection pursuit regression. The conclusions made here must also be judged in light of the fact that it was based on two specific data sets which is probably not representative of all NIR data sets. The option of simulating a wider variety of data sets will therefore be investigated.

To summarise, these initial results indicate that MARS, and to a lesser degree SVM, are superior techniques to PLS, which is currently the preferred technique for NIR calibration.

References

Frank, I.E. and Friedman, J.H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**(2), 109–135.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference and Prediction.* Springer Series in Statistics.

Seasholtz, M.B. and Kowalski, B. (1993). The parsimony principle applied to multivariate calibration. *Analytica Chimica Acta*, **277**, 165–177.

Sjöström M. and Wold, S. (1983). A multivariate calibration problem in analytical chemistry solved by partial least-squares models in latent variables. *Analytica Chimica Acta*, **150**, 61–70.

Simultaneous Regression modelling of Means and Correlations in Lung Function for Spouses: an Application of the FISHER Software.

Matthew W. Knuiman[1] and Mark L. Divitini[1]

[1] School of Population Health, University of Western Australia, Crawley WA 6009, Australia

Abstract: Correlations between spouses in health-related variables are determined by partner selection factors and shared environmental factors during marriage. It is believed that spousal correlations are positive and increase with duration of marriage. Study of spouse correlations in quantitative variables requires simultaneous regression modeling of means and correlations. The program FISHER was used to model spousal correlations in lung function measures from the Busselton Health Study [http://bsn.uwa.edu.au]. The model for the mean included effects of gender, age, height and smoking, and various models for the correlation in relation to duration of marriage and age at marriage were explored. The models and results indicated that (a) adjustment in the mean model for covariates that are key determinants of lung function and which themselves have considerable spouse concordance substantially reduces the estimated spousal correlation in lung function, (b) there is clear evidence of non-random partner selection in relation to lung function, (c) spousal correlation in lung function does not appear to vary with duration of marriage but does decline with age at marriage.

Keywords: FISHER software; Spousal correlations; Regression models; Multivariate normal distribution; Lung function.

1 Motivation

There is considerable interest in identifying genetic, lifestyle and environmental factors that contribute to disease development (Palmer et al 2001). Family studies, based on married couples and their offspring are used to identify and separate genetic, lifestyle and environmental factors. Studies of (non-genetically related) married couples are useful in understanding environmental factors associated with cohabitation or the sharing of a common environment. Married couples often share lifestyle factors (such as smoking) as well as a common household and local environment (Venters et al 1984). Concordance between spouses in lifestyle and environmentally-related risk factors is due to partner selection factors ('assortative mating') and the effects of marriage/cohabitation on lifestyle and environmental factors. Both

groups of factors are believed to induce positive concordance/correlation in risk factors for disease and the concordance/correlation is expected to increase with marriage duration.

2 Statistical Models and FISHER

Study of spouse correlations in (quantitative) disease-related variables in relation to lifestyle factors and marriage duration requires statistical models that allow for simultaneous regression modelling of means and correlations. The program FISHER is freely available and allows flexible modelling of means, variances, covariances and correlations for multivariate normal data (Lange et al, 1976, 1988; Hopper 1993; Hopper et al, 1994).

3 Busselton Lung Function Data

Busselton is a town in Western Australia and its residents have been the subject of several health surveys over the period 1966 to 1995. Surveys of adults in Busselton were conducted in 1966, 1969, 1972, 1975, 1978, 1981 and 1994/95 [http://bsn.uwa.edu.au]. Lung function was measured by spirometry. FEV1 (forced expiratory volume in 1 second) is the exhaled volume in first second and is a measure of breathing capacity. FVC (forced vital capacity) is the total exhaled volume and their ratio FEV1/FVC is a measure of airway narrowing. Sex, age, height and smoking are key determinants of lung function. A total of 2,617 husband-wife pairs attended at least one survey together since 1969 (lung function data were not collected for women in 1966). For this analysis, lung function data were taken from the survey attended when marriage duration was smallest to try to get more data on the first 5-10 years of marriage. The focus of the analysis is on the spousal (ie husband-wife) correlation in lung function measures overall and in relation to duration of marriage and age at marriage. The average age (at lung function measurement) of husbands and wives were 46.7 and 43.2 years respectively, the average duration of marriage was 17.2 years, and hence the average age at marriage was 29.5 years for husbands and 26.0 years for wives. The overall correlations between husbands and wives were 0.955 for age, 0.268 for height, 0.267 for smoking (coded as 1 = never, 2 = former, 3 = current), 0.540 for FEV1 and 0.367 for FEV1/FVC.

4 Regression Models for Spouses

A bivariate normal model for lung function in spouse pairs was used. The models for the mean included (progressively) the effects of gender, age (gender-specific quadratic trends), height (gender specific linear trends), and smoking (gender specific effects for never, former and current). The

variance was assumed constant and various models for the correlation in relation to duration of marriage and age at marriage were explored, including constant, grouped and linear trend models.

5 Main Results

The estimated residual variance in FEV1 declined from 0.67 to 0.26 when the model for the mean progressively included terms representing the effects of age, height and smoking, confirming that these are key determinants. The residual variance for FEV1/FVC % similarly declined from 105 to 75. The estimated husband-wife correlation in FEV1 declined from 0.54 to 0.08 when the model for the mean progressively included terms representing the effects of age, height and smoking. The FEV1/FVC correlation similarly declined from 0.37 to 0.14. The adjustment for age had the greatest effect on the estimated residual variance and correlation. Estimated spousal correlations in lung function by marriage duration groups indicated little trend in the correlations with marriage duration. The correlation models that included linear trends with marriage duration and age at marriage confirmed non-significant trends with marriage duration but revealed a declining trend for age at marriage ($p = 0.004$ for FEV1 and $p < 0.001$ for FEV1/FVC).

6 Comments on Statistical Issues

The program FISHER is cumbersome to use but does allow flexible simultaneous modelling of means and correlations. As expected, adjustment in the mean model for key determinants of lung function substantially reduces the residual variance. Adjustment in the mean model for covariates that are key determinants of lung function AND which themselves have considerable positive spouse concordance substantially reduces the spousal correlation in lung function. Estimated trends (with marriage duration or age at marriage) were not influenced by degree of adjustment for key determinants in mean model. This analysis of trends based on cross-sectional data is open to possible biases. For example, bias may be introduced if marriages for couples initially discordant for lung function are more likely to terminate due to divorce or death of one partner. Longitudinal studies that repeatedly measure (throughout married life) cohorts of newly married couples are required.

7 Comments on Respiratory Epidemiology Issues

There is clear evidence of non-random partner selection in relation to lung function measures and most (but not all of it) is explained by age and

height. Spousal correlation in lung function does not appear to vary with marriage duration. Spousal correlation in lung function appears to vary (actually decline) with age at marriage. Spousal concordance in lung function measures appears to be dominated by partner selection factors and age at marriage and common exposure to lifestyle/household/neighbourhood influences has little effect. This should be recognised and considered in family studies that aim to identify and separate genetic from other influences.

Acknowledgments: The authors thank the community of Busselton for their long standing support for the Busselton Health Study, the Busselton Population Medical Research Foundation for access to the data, and Ms Helen Bartholomew for database programming support. This work was supported by grant number 211988 from the National Health and Medical Research Council of Australia.

References

Hopper, J.L. (1993). Variance components for statistical genetics: applications in medical research to characteristics related to human disease and health. *Statistical Methods in Medical Research*, **2**, 199-223.

Hopper, J.L. and Matthews, J.D. (1994). A multivariate normal model for pedigree and longitudinal data and the software FISHER. *Australian Journal of Statistics*, **36**, 153-176.

Lange, K., Westlake, J., and Spence, M.A. (1976). Extensions to pedigree analysis III: Variance components by the scoring methods. *Annals of Human Genetics*, **39**, 485-491.

Lange, K., Weeks, D., and Boehnke. M. (1988). Programs for pedigree analysis: MENDEL, FISHER and dGENE. *Genetic Epidemiology*, **5**, 471-472.

Palmer, L.J., Knuiman, M.W., Divitini, M.L., Burton, P.R., James, A.L., Bartholomew, H.C., Ryan, G., and Musk, A.W. (2001). Familial aggregation and heritability of adult lung function: results from the Busselton Health Study. *European Respiratory Journal*, **17**, 696-702.

Venters, M.H., Jacobs, D.R., Luepker, R.V., Maiman, L.A., and Gillum, R.F. (1984). Spouse concordance of smoking patterns: the Minnesota Heart Survey. *American Journal of Epidemiology*, **120**, 608-616.

Accelerated Failure Time Model for Arbitrarily Censored Data with Smoothed Error Distribution

Arnošt Komárek[1], Emmanuel Lesaffre[1], and Joan F. Hilton[2]

[1] Catholic University Leuven, Biostatistical Centre, Kapucijnenvoer 35, B–3000 Leuven, Belgium

[2] University of California San Francisco, Dept. of Epidemiology and Biostatistics, 500 Parnassus Avenue, MU–420W, Box 0560, CA 94143–0560 San Francisco, USA

Abstract: In this article we develop a procedure to estimate parameters in the accelerated failure time model whose error distribution does not have to be specified and it is estimated using the smoothing techniques. First, a density of the error distribution is specified as a linear combination of P-splines (B-splines with penalties), see Eilers and Marx (1996) and second, B-splines are replaced by their limits which appear to be Gaussian densities. We call the resulting smoothed function as a G-spline. The spline coefficients as well as the regression parameters are estimated via a constrained penalized maximum likelihood method. The procedure allows for all types of censoring (left, right and interval). The method is illustrated on the analysis of the dataset from AIDS research.

Keywords: Accelerated failure time model; B-splines; Penalized likelihood.

1 Introduction

The accelerated failure time model (AFT) is a worthwhile alternative to the Cox's proportional hazards model. This model specifies that the effect of a vector of fixed covariates $\mathbf{x}$ acts additively on the logarithm of the time to event T as

$$\log(T) = Y = \alpha + \beta^T \mathbf{x} + \sigma\varepsilon, \tag{1}$$

where ε is the error term with a density $f(e)$, α and β are regression parameters and σ is a scale parameter. The expression (1) is simply a linear model on the log scale of time but unlike the area of uncensored data where the normal distribution is the most used error distribution, there is no gold standard distribution for censored data. Moreover, in survival analysis non- or semi-parametric procedures are generally preferred.

2 B- and P-splines

Our approach assumes that the density $f(e)$ of the error term can be well approximated by a mixture of B-splines (de Boor, 1978). The density of the error term is first assumed to have a form of a spline function of degree k defined on a finite interval (e_{min}, e_{max}). Briefly, a spline function of degree k is formed of polynomial pieces of the same degree and all derivatives up to order $k-1$ are continuous. Polynomial pieces are connected together at so called knots $\eta_{-k} = \ldots = \eta_0 = e_{min} < \eta_1 < \cdots < \eta_{g^*} < e_{max} = \eta_{g^*+1} = \ldots = \eta_{g^*+k+1}$. The spline function (density of the error term) can then be represented as a mixture of basis B-splines $N_{i,k+1}$ of degree k as follows $f(e|\mathbf{c}) = \sum_{i=-k}^{g^*} c_i N_{i,k+1}(e)$. The basis B-spline $N_{i,k+1}$ is defined through the knots $\eta_i, \ldots, \eta_{i_k+1}$, is positive on (η_i, η_{i+k+1}) and zero elsewhere. Restrictions on c_i's such that $0 < c_i < 1$ and $\sum_{i=-k}^{g^*} c_i = (e_{max} - e_{min})^{-1}$ ensure that the resulting spline function is a density.
Choosing the optimal number and the position of knots is generally a complex task. Too many knots lead to over fitting of the data, too few knots lead to under fitting. O'Sullivan (1988) proposed to use a relatively large number of knots and to restrict the flexibility of the fitted curve by putting a penalty on the second derivative. Eilers and Marx (1996) further generalized this approach in the context of B-splines. Basically, the penalized log-likelihood is maximized for computing the estimates of the parameters. The knots can be chosen equidistantly and there is no need to search for an optimal number. Eilers and Marx use consequently the term P-splines instead of B-splines to stress the fact that the spline coefficients are estimated via the maximization of the penalized log-likelihood.

3 G-splines

A possible drawback of the above described spline approach could be the finite support (e_{min}, e_{max}) of the fitted density function. However, the basis B–spline of degree k is proportional to the density function of a sum of $k+1$ independent uniformly distributed random variables. One can show then that after a proper normalization including an expansion of the basis B–spline support it converges uniformly $(k \to \infty)$ on $\Re$ to a Gaussian density, see Unser et al. (1992) for details. That is why we concluded that a mixture of Gaussian densities could be used as a model for the error density instead of the original B-splines mixture. Thus, the density of the error distribution can be now represented as

$$f(e|\mathbf{c}) = \sum_{j=1}^{g} c_j \varphi_{\mu_j,\sigma_0}(e) \tag{2}$$

where φ_{μ_j,σ_0} stands for a density of $\mathrm{N}(\mu_j, \sigma_0^2)$. To get a proper density function constraints $\sum_{j=1}^{g} c_j = 1, c_j > 0, j = 1, \ldots, g$ are imposed on

coefficients c_j. To keep the correspondence to the original spline approach we will call each part of the linear combination (2), i.e. a density of $\mathrm{N}(\mu_j, \sigma_0^2)$ as a G-spline (G standing for Gaussian). Means $\mu_1 < \ldots < \mu_g$ are now used instead of original knots and are fixed during the computation of estimates as well as the variance σ_0^2 of each basis G-spline. Means μ_j will be still called as knots in the following.
The G-spline coefficients c_j will be estimated using the penalized maximum likelihood method as will be explained below. For this purpose we use a grid of equidistant knots. According to our experience (supported by simulations) the distance of 0.3 between the two consecutive knots is usually sufficient when smoothing a standardized (zero mean, unit variance) density.

4 Constrained Penalized Maximum Likelihood Method

All parameters in the model, i.e. spline coefficients $\mathbf{c} = (c_1, \ldots, c_g)^T$, regression parameters α, β and the scale σ will be estimated by the mean of the constrained penalized maximum likelihood method. Earlier mentioned constraints $\sum_{j=1}^{g} c_j = 1, c_j > 0, j = 1, \ldots, g$ can be omitted through a reparametrization of the problem as $c_j(\mathbf{a}) = e^{a_j}(\sum_{l=1}^{g} e^{a_l})^{-1}$ with one of the new a_j coefficients fixed to a particular value. We may assume without loss of generality $a_1 = 0$.
To clearly distinguish the regression part of the model (1) from the error distribution (2) we fix the mean and the variance of the fitted error distribution to zero and one, respectively. That is we impose the following constraints on the $\mathbf{a}$ parameters, $0 = \mathrm{E}(\varepsilon) = \sum_{j=1}^{g} c_j(\mathbf{a})\mu_j$, and $1 = \mathrm{var}(\varepsilon) = \sum_{j=1}^{g} c_j(\mathbf{a})(\mu_j^2 + \sigma_0^2)$. Subsequently, a penalized log-likelihood $\ell_{P,n}(\theta|\lambda; \mathbf{y}) = \ell_n(\theta|\mathbf{y}) - \frac{\lambda}{2}\sum_{j=m+1}^{g}(\Delta^m a_j)^2$ is maximized w.r.t. θ under the two constraints. The vector $\mathbf{y} = (y_1, \ldots, y_n)^T$ denotes a set of n independent (possibly censored) responses, and $\theta = (\alpha, \beta^T, \log(\sigma), \mathbf{a}^T)^T$ is a vector of unknown parameters to be estimated with $\mathbf{a} = (a_2, \ldots, a_g)^T$. The operator Δ^m stands for the ordered difference, i.e. $\Delta^1 a_j = a_j - a_{j-1}$, $\Delta^{m+1} a_j = \Delta^m a_j - \Delta^m a_{j-1}$. The parameter λ is a tuning parameter which determines a degree of smoothing. The term $\ell_n(\theta|\mathbf{y})$ is an ordinary log-likelihood based on possibly censored responses $\mathbf{y}$ under the model (1) with the error density (2).
The choice of the tuning parameter can be based on the Akaike's information criterion $AIC(\lambda) = \ell_{P,n}(\hat{\theta}(\lambda)|\lambda; \mathbf{y}) - df(\lambda)$ where $df(\lambda)$ is the effective number of parameters defined in the similar manner as in Gray (1992). The effective number of parameters varies between $\dim(\beta) + 1$ for $\lambda \to \infty$ and $\dim(\beta) + g - 1$ for $\lambda = 0$ and describes the estimated number of parameters while adjusting for a degree of the penalization.

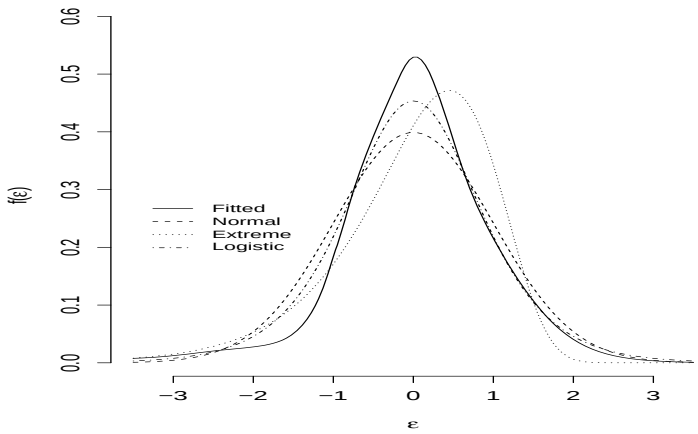


FIGURE 1. *MHC Study. Fitted error distribution compared to the standardized densities of the normal, extreme value and logistic distribution.*

Using similar technique as in O'Sullivan (1988), modified appropriately to take into account the two constraints a Bayesian technique for generating confidence bands for penalized MLE's can be used.
The approach described has been implemented as a function in R and the function is available upon request from the first author. The simulation study was performed to evaluate the characteristics of the suggested approach. The error term in the simulation study was sampled from normal, extreme value distribution and from the mixture of the two normals and satisfactory results were obtained.

5 Multicenter Hemophilia Cohort Study

Engels et al. (1999) evaluated the relation between plasma HIV viral load and subsequent risk for disease progression in patients with hemophilia and late-stage HIV disease (CD4 count < 200 cells/mm^3) using a subset of the Multicenter Hemophilia Cohort Study (MHC Study). See Goedert et al. (1989) for more details on the setup of the study. They used various Cox's PH–models and stratified Kaplan–Meier estimates without accounting for interval censoring of the response (development of clinical AIDS).
It can be thus interesting to try models which do account for interval censoring since ignoring it may introduce a bias. We use a subgroup of 335 hemophilic HIV positive men/boys who were between 2.5 and 30 years old at the baseline visit. Our sample corresponds very closely but not exactly with the sample analyzed by Engles et al. Values of the plasma HIV viral load, CD4 counts and CD8 counts at the baseline visit are available.

TABLE 1. *MHC Study. Estimates in the AFT model with smoothed error compared to the estimates in the AFT model with extreme value error.*

Parameter	Smoothed error	Extreme value error
$\hat{\beta}(CD4)$	$7.99\ (2.53)\ \cdot 10^{-4}$	$7.82\ (2.38)\ \cdot 10^{-4}$
$\hat{\beta}(vl4)$	$-0.49\ (0.20)$	$-0.56\ (0.22)$
$\hat{\beta}(vl5)$	$-0.92\ (0.20)$	$-1.01\ (0.23)$
$\hat{\beta}(vl6)$	$-1.18\ (0.23)$	$-1.09\ (0.24)$
$\hat{\beta}(CD8)$	$-0.76\ (1.23)\ \cdot 10^{-4}$	$-1.12\ (1.31)\ \cdot 10^{-4}$
$\hat{\beta}(age)$	$4.21\ (11.64)\ \cdot 10^{-3}$	$2.56\ (11.48)\ \cdot 10^{-3}$
Log–likelihood	-419.8	-425.0

We fitted models with time to the onset of clinical AIDS in months as the response which was interval censored with mean length of the intervals equal to 10 months. In the AFT model with CD4, CD8 counts, age at baseline and dummies for intervals of viral load at the baseline defined as less than 10^4, $[10^4, 10^5)$ $(vl4)$, $[10^5, 10^6)$ $(vl5)$, and at least 10^6 copies/mL $(vl6)$, only the baseline viral load and CD4 at baseline appeared to be significant (5%). The fitted error distribution compared to three other widely used error densities is shown on Figure 1. The AIC of the fitted model was minimized for $\lambda = 2$ and took the value of -431.3 Models with specified error distributions (normal, extreme value or logistic respectively) gave similar estimates of the regression parameters as our procedure. Comparison of our estimates to the estimates in the extreme value model which showed the highest likelihood among the parametric models is shown in Table 1.
Although, only a qualitative comparison to the results of Engels et al. is possible our conclusions are similar to these drawn by Engels et al. who concluded that each $\log_{10}$ increase in baseline viral load was associated with rather high increase in risk for AIDS–related illness during the first few months of the follow–up. Based on our model, the expected time to the development of AIDS–related illness, after adjusting for the CD4 count is 0.6, 0.4 or 0.3 times respectively lower than the expected time for the person with less than 10 000 copies of the virus/mL if the viral load is 10, 100 or 1 000 times respectively higher.

Acknowledgments: The first two authors acknowledge support from the Interuniversity Attraction Poles Program P5/24 – Belgian State – Federal Office for Scientific, Technical and Cultural Affairs. The authors thank James J. Goedert, M.D., Division of Cancer Epidemiology and Genetics, National Cancer Institute, for sharing the MHCS data. The research of the third author was funded by the National Institute for Dental and Cranio-

facial Research, P01–DE07946.

References

de Boor, C. (1978). *A Practical Guide to Splines.* Springer, Berlin.

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

Engels, E.A., Rosenberg, P. S., O'Brien, T. R., and Goedert, J. J. (1999). Plasma HIV viral load in patients with hemophilia and late-stage HIV disease: A measure of current immune suppression. *Annals of Internal Medicine,* **131,** 256–264.

Goedert, J.J., Kessler, C.M., and Aledort, L.M. (1989). A prospective study of human immunodeficiency virus type I infection and the development of AIDS in subjects with hemophilia. *The New England Journal of Medicine,* **321,** 1141–1148.

Gray, R.J. (1992). Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *Journal of the American Statistical Association,* **87**, 942–951.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal of Scientific Computing,* **9,** 363–379.

Unser, M., Aldroubi, A., and Eden, M. (1992). On the asymptotic convergence of B-spline wavelets to Gabor functions. *IEEE Transactions on Information Theory,* **38,** 864–872.

Gaussian Modelling of Non-Gaussian Time Series

D. Kugiumtzis[1] and E. Bora-Senta[2]

[1] Department of Mathematical, Physical and Computational Sciences, Faculty of Engineering, Aristotle University of Thessaloniki, Thessaloniki 54006, Greece
[2] Department of Mathematics, Faculty of Mathematics and Natural Science, Aristotle University of Thessaloniki, Thessaloniki 54006, Greece

Abstract: A framework that allows the use of Gaussian linear analysis on non-Gaussian time series is proposed. The idea is to approximate first the transform that renders the marginal distribution Gaussian, and from this transform to determine the autocorrelation of the Gaussian time series as a function of the original one. The approximation of the transform is chosen to be piecewise polynomial and the moments of the truncated normal distribution are used to determine the relationship for the autocorrelations. The derived Gaussian time series has the property that through the inverse transform it possesses the same linear correlations and marginal distribution as the original time series. Thus the standard linear analysis and modeling can be performed on this Gaussian time series and the results of the analysis, passed through the inverse transform, can yield the original non-Gaussian time series. This approach is particularly useful for the surrogate data test for nonlinearity which relies heavily on the generation of proper surrogate time series that possess the linear correlations and marginal distribution of a given time series. The importance of this approach both for the linear modeling of time series and for the surrogate data test for the nonlinearity will be illustrated with some real world time series from finance and physiology.

Keywords: Time series; Non-Gaussian; Nonlinearity; Surrogate data test.

1 Introduction

The linear analysis of time series is well established, especially for Gaussian time series, where, for example, best predictors are linear and the statistics of the optimal estimators can be computed analytically (Brockwell and Davis, 1991; Rosenblatt, 2000). If the time series cannot be assumed Gaussian, statistics pertaining the fitted model can be computed numerically through bootstrapping techniques. In some cases, simple transforms, such as logarithms, may render the Gaussian marginal distribution and allow analytical results. One should be careful though when transforming time series because a Gaussian marginal distribution does not necessarily imply a Gaussian generating process. Furthermore, the linear correlations

may be altered through the transform and this should be taken care of in the analysis of the transformed time series.
For a given non-Gaussian time series we attempt to generate a Gaussian time series with suitable autocorrelation, so that under a static transform it has the same marginal distribution and autocorrelation as the given time series. Thus the analysis and modeling can be performed on this Gaussian time series and the results, such as point predictions, confidence and prediction intervals, can be mapped through the static transform to yield the original time series.
The motivation for this approach stems from the surrogate data test for nonlinearity and in particular from the problem of generating time series that possess the linear correlations and amplitude distribution of a given time series (Theiler et al, 1992; Schreiber and Schmitz, 2000; Kugiumtzis, 2002a). Such time series are called "surrogate data" and are used to represent the null hypothesis that a given time series is stochastic linear (explicitly, the time series is generated by a Gaussian process possibly undergoing a static transform, linear or nonlinear). The generation of surrogate data using a simple polynomial approximation for the transform of the marginals was recently used to improve significantly the performance of the test (Kugiumtzis, 2002b).
In this paper, we extend the approximation of the transform to piecewise polynomial in an attempt to provide better estimation for the linear correlations. The accurate estimation of the linear correlations is not only useful for the surrogate data test for nonlinearity, but also for the linear modeling of non-Gaussian time series. The enhanced approach generates a Gaussian time series, which is linearly equivalent to the original non-Gaussian time series, so that the modeling is done on this Gaussian time series and the estimates and predictions can be transformed back to the original time series. In the following, we draw the main points of the approach.

2 Gaussian from Non-Gaussian Time Series

Let us suppose a transform g that maps two variables s_1 and s_2 having standard Gaussian joint distribution to the variables $x_1 = g(s_1)$ and $x_2 = g(s_2)$. For some known joint distributions of x_1, x_2, analytic expressions exist for the transform ψ for the corresponding correlation coefficients ρ_s and ρ_x, such that $\rho_x = \psi(\rho_s)$ (Hutchinson and Lai, 2001). This result can be extended to time series, i.e. for two time series $\{s_i\}$ and $\{x_i\}$ where $x_i = g(s_i)$, there exists a function ψ, such that $\rho_x(\tau) = \psi(\rho_s(\tau))$, where $\rho(\tau)$ is the autocorrelation for delay τ. For an arbitrary distribution of x_1, x_2 (and subsequently of $\{x_i\}$) the transform g can be expressed by the rank ordering

$$x_i = g(s_i) = \Phi_x^{-1}(\Phi_0(s_i)), \tag{1}$$

where Φ_x is the marginal cumulative density function (cdf) for x and Φ_0 is the standard Gaussian cdf. The transform g as defined above is monotonic and the inverse transform g^{-1} is defined in a similar way. However, to the best of our knowledge, an analytic expression of ψ for the correlation coefficients (and subsequently for the autocorrelations) does not exist in general.

In a first approach in (Kugiumtzis 2002b), the transform g is approximated by a polynomial of some degree m. For a given time series $\{x_i\}$, the polynomial is estimated from the graph of $\{x_i\}$ versus generated standard normal white noise data reordered to match the rank order of $\{x_i\}$. It was found that then ψ is also a polynomial of degree m with coefficients $c_i, i = 1, \ldots, m$, given in terms of the coefficients of the polynomial for g

$$\rho_x = \psi(\rho_s) = \sum_{i=1}^{m} c_i \rho_s^i. \tag{2}$$

The correlation coefficient ρ_s can be found from the solution of eq(2). In (Kugiumtzis, 2002b), it is conjectured that a unique solution exists but still there is no theoretical proof for this. For time series, eq(2) can be solved for each τ substituting ρ_x by $\rho_x(\tau)$ and ρ_s by $\rho_s(\tau)$ in order to derive the autocorrelation function ρ_s for the desired Gaussian time series. Indeed, ρ_s alone defines a standard Gaussian process which under the transform g, as defined in eq(1), generates time series that have marginal cdf Φ_x and autocorrelation ρ_x.

This approach has been used to generate surrogate data possessing F_x and r_x (the sample cdf and autocorrelation) of a given time series $\{x_i\}, i = 1, \ldots, n$. The complete algorithm of statically transformed autoregressive process (STAP) provides proper surrogate time series $\{z_i\}$: $F_z(z_i) = F_x(x_i)$ is attained exactly and r_z is an unbiased estimate of r_x (Kugiumtzis, 2002b). Compared to the two most known algorithms for surrogate data generation, the amplitude adjusted Fourier transform (AAFT) (Theiler et al., 1992) and the iterated AAFT (IAAFT) (Schreiber and Schmitz, 1996), the test for nonlinearity turned out to perform best with STAP.

In the proposed paper the polynomial approximation of the transform g is extended to piecewise polynomial to reach better fit. The variable s_i, restricted at each interval of the partition, follows a truncated normal distribution. The moments of the joint truncated normal distribution for s_1, s_2 are expressed in terms of the truncation points, which are known from the selected partition, and the correlation coefficient ρ_s (Johnson and Kotz, 1990; Regier and Hamdan, 1971). Making use of the expressions for the moments we determine ψ that gives ρ_x in terms of ρ_s as in eq(2), but here the expressions of the coefficients c_i are rather involved. Thus the procedure for generating the equivalent Gaussian time series is the same as for the simple polynomial approximation of g, but the estimation of ρ_s is improved to the cost of more intensive computations.

3 Work in Progress

We are now working on testing with simulated time series for the correctness of the estimation of ψ and thus ρ_s. Then we intend to build an algorithm similar to STAP and study the improvement of the accuracy in the match of linear correlations with the use of piecewise polynomials. Finally, this approach will be used to make predictions of non-Gaussian time series using the following procedure. A linear model is estimated from the autocorrelation ρ_s of the equivalent Gaussian time series and the predictions are transformed by g to derive the predictions for the original time series. For this paper, we plan to illustrate the performance of the enhanced approach with the same real world time series as those used in conjunction with the STAP algorithm in (Kugiumtzis 2002a; Kugiumtzis 2002b), i.e. electroencephalographs (EEG) from normal and epileptic activity and volatility data from the exchange rates of USD/GBP.

References

Brockwell, P.J. and Davis, R.A. (1991). *Time Series: Theory and Methods*. Springer Series in Statistics. New York: Springer-Verlag.

Rosenblatt, M. (2000). *Gaussian and Non-Gaussian Linear Time Series and Random Fields*. New York: Springer-Verlag.

Theiler, T., Eubank, S., Longtin, A., and Galdrikian, B. (1992). Testing for nonlinearity in time series: the method of surrogate data. *Physica D*, **58**, 77-94.

Schreiber, T. and Schmitz, A. (2000). Surrogate time series. *Physica D*, **142**, 346-382.

Kugiumtzis, D. (2002a). Surrogate data test on time series. In *Modelling and Forecasting Financial Data, Techniques of Nonlinear Dynamics*, 267-282. Kluwer Academic Publishers.

Kugiumtzis, D. (2002b). Statically transformed autoregressive process and surrogate data test for nonlinearity. *Physical Review E*, **66**, 025201.

Hutchinson, T.P. and Lai, C.D. (1990). *Continuous Bivariate Distributions Emphasising Applications*. Rumsby Scientific Publishing.

Schreiber, T. and Schmitz, A. (1996). Improved surrogate data for nonlinearity tests. *Physical Review Letters*, **77**, 635-638.

Johnson, N. and Kotz, S. (1990). *Distributions in Statistics, Continuous Multivariate Distributions*. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

Regier, M.H. and Hamdan, M.A. (1971). Correlation in a bivariate normal distribution with truncation in both variables. *Australian Journal of Statistics*, **13**, 77-82.

Modelling Physicians' Recommendations for Optimal Medical Care by Random Effects Stereotype Regression

Oliver Kuss[1]

[1] Institute of Medical Epidemiology, Biostatistics, and Informatics, University of Halle-Wittenberg, D-06097 Halle (Saale), Germany

Abstract: We show how Anderson's Stereotype regression model can be extended to account for correlated responses by a simple nonlinear parameter restriction on the multinomial logistic model with random effects. A data set on physicians' recommendations and preferences in traumatic brain injury rehabilitation is used for illustration.

Keywords: Stereotype regression; Random effects; Multinomial logistic regression; Traumatic brain injury.

1 Introduction

Many study designs in applied sciences give rise to correlated data. For example, subjects are followed over time, are repeatedly treated under different experimental conditions, or are observed in logical units (e.g. clinics, families, litters).
One of the standard analysing tools in these situations which adequately accounts for the correlation between observations is the random effects (RE) model, sometimes also called hierarchical or mixed effects model. This is quite common for continuous responses, and also, despite an enhanced mathematical complexity, for binary responses. Less used have been random effect models for the analysis of discrete non-binary responses, some of the rare examples are Hedeker and Gibbons (1994) and Tutz and Hennevogl (1996) for ordinal and Hartzel et al (2001) for nominal responses. To our knowledge, up to now there exists no random effects version of the Stereotype regression model.
The Stereotype regression model was originally proposed by Anderson (1984). He observed that some relevant discrete non-binary responses in applied statistics are not perfectly ordinal in the sense that there is an latent continuous variable which was only observed in discrete and disjunct classes, but should rather be regarded as a multidimensional phenomenon where several items determine the grade on the ordinal scale, the most prominent example being maybe the severity of a disease.

2 The Random Effects Stereotype Regression Model

To extend the Stereotype regression model to account for correlated responses we use the fact that the original Stereotype model is derived from the ordinary multinomial logistic regression model by a certain nonlinear parameter restriction. This restriction is simply applied to the multinomial logistic random effects model of Hartzel et al (2001). The Stereotype model with random effects thus becomes a nonlinear model with random effects and all the well-known theory and estimation methods (see e.g. Davidian and Giltinan, 1995) can be used.

We assume that our data comprises a set of I $(i = 1, \ldots, I)$ independent clusters where the i-th cluster consists of n_i observations. Let Y_{ij} denote the j-th response in cluster i $(j = 1, \ldots, n_i)$, where this response is from one of r $(r = 1, \ldots, R)$ distinct categories and the response probability is $\pi_{ijr} = P(Y_{ij} = r)$. Further, x_{ij} denotes a column vector of covariates for the j-th observation in the i-th cluster. Thus the model equation is

$$\log\left(\frac{\pi_{ijr}}{\pi_{ijR}}\right) = \theta_r + x'_{ij}\phi_r\beta + u_{ir}, \quad r = 1, \ldots, R-1, \tag{1}$$

where the θ_r are constant terms, the scalars ϕ_r introduce a metric for the common effect of the covariates, where this effect is assumed constant across response categories. The influences of covariates are assessed through the components of $\beta = (\beta_1, \ldots, \beta_p)'$. The θ_r, the ϕ_r, and the β are considered to be fixed effects. For the random effects u_{ir} we assume a multivariate normal distribution with unstructured covariance matrix Σ, that is for $u_i = (u_{i1}, \ldots, u_{i,R-1})'$ we have $u_i \sim N(0, \Sigma)$.

For reasons of identification of parameters we restrict $\theta_R = 0$, $\beta_R = 0$, $u_R = 0$, $\phi_1 = 0$, and $\phi_R = 1$, so that interpretation of parameters is, analogous to the multinomial logistic model, with reference to the R-th category. Note that the model equation of the RE Stereotype regression model is derived from the multinomial logistic random effects model of Hartzel et al (2001) by the non-linear parameter restriction $\beta_r = \phi_r\beta$.

The estimation of parameters is complicated by the fact that the likelihood function consists of a product of I integrals which can not be solved in closed form. Thus, numerical or stochastic integration are viable alternatives. Hartzel et al (2001) suggest adaptive Gaussian quadrature as the preferred method for parameter estimation in this model class. As such, the model can be fitted conveniently with, for example, SAS PROC NLMIXED.

3 The Motivating Example

The motivation for the derivation of the RE Stereotype model was a data set from a study on physicians' recommendations and preferences in traumatic brain injury (TBI) rehabilitation (Hasenbein et al, 2003). In this

study, 36 physicians were asked to decide on the optimal rehabilitation setting (in-patient, day-clinic, out-patient) for each of ten typical TBI disease histories. Of course, we expect the setting recommendations within the same physician to be correlated. Concerning the 3-valued response we recognize that this is not strictly nominal, but has indeed some ordinal flavor, for example, we might think of the "time not at home" as some underlying continuous variable. However, it is not that simple that in-patient, day-clinic, and out-patient rehabilitation only differ by the time that patients stay in the clinic, instead they rather represent different therapeutic concepts and actual treatment varies. Of interest was mainly if we could identify factors (considering physicians and disease histories) that influence setting preferences.
In the following (see Table 1) we give the results (estimates and respective standard errors in parentheses) for our data set for the ordinary Stereotype model, the RE multinomial model and the RE Stereotype model. Four covariates, all of them binary, were included in the model, two of them referring to physicians' characteristics (1. Is the physician a neurologist [NEURO] and 2. Is the physician a specialist [SPECIAL]) and two describing the disease history (3. Is the time since the event longer than 3 months [TIME] and 4. Is the patient severely or moderately handicapped after the TBI [SEVERITY]). As the reference category of the response we chose the stationary setting, and compare day-clinic (DC) and out-patient (OP) to this.
Some remarks regarding the results can be made: As we expect (and maybe hope as potential patients), physicians' own characteristics do have only small influence on their recommendations. Looking at the values of the model selection criteria we see that the random effects Stereotype model is superior to the other two models: Compared to the ordinary Stereotype model this means on one hand that it is essential to account for the inherent correlation in the data (which is also confirmed by the significant values of the random effects covariance matrix). Compared to the random effects multinomial model on the other hand we note that we do not need the additional information of looking separately at the two response categories, instead the RE Stereotype model gives a natural summary of the ordering of response categories and judges the DC category roughly in the middle ($\phi_2 = 0.55$) between the reference category and the OP category. Summing up a bit roughly in subject matters: The more severe the TBI and the shorter the time since TBI, the more time the patient should spend in the hospital.

4 Discussion

We showed how Anderson's Stereotype regression model can be extended easily to account for correlated responses. The idea was to impose the nonlinear parameter restriction which relates the ordinary multinomial logistic

TABLE 1. *Results (estimates and respective standard errors in parentheses) from the ordinary Stereotype model, the RE multinomial model and the RE Stereotype model for the TBI data set*

	Stereotype Model	RE Multinomial Model		RE Stereotype Model
		Fixed effects		
		DC	OP	
$\hat{\beta}_{NEURO}$	0.89 (0.46)	-0.56 (0.74)	-0.36(0.90)	1.36 (0.93)
$\hat{\beta}_{SPECIAL}$	0.19 (0.43)	-0.63 (0.78)	-0.04 (0.94)	0.40 (0.88)
$\hat{\beta}_{TIME}$	3.26 (0.45)	2.51 (0.41)	3.43 (0.50)	4.23 (0.58)
$\hat{\beta}_{SEVERITY}$	-2.00 (0.43)	-1.94 (0.43)	-3.29 (0.47)	-2.60 (0.53)
$\hat{\phi}_2$	0.50 (0.10)	–	–	0.55 (0.09)
		Random effects		
$\hat{\sigma}_1^2$	–	1.47 (0.35)	–	2.01 (0.96)
$\hat{\sigma}_2^2$	–	–	1.87 (0.43)	2.62 (1.20)
$\hat{\sigma}_{12}^2$	–	2.54 (1.11)		1.94 (0.93)
		Model selection criteria		
AIC	517.3	499.3		487.5
BIC	543.9	516.7		503.3

model to the original Stereotype model to the random effects multinomial model of Hartzel et al (2001). Proceeding that way, the RE Stereotype model becomes a nonlinear random effects model and standard theory and estimation methods apply. In terms of our motivating example we were able to identify factors which influence physicians' preferences on optimal rehabilitation setting in TBI patients. We learned that we had to account for the inherent correlation in the data but did not need the additional complexity of the RE multinomial model. Moreover, we got information about distances between response categories. The estimation method of numerical integration seems to work well as some limited preliminary evidence from simulation studies reveals. In the future we are mainly interested in additional estimation techniques to judge robustness of our results, where MCMC and nonparametric ML methods might be promising candidates.

Acknowledgments: We are grateful to Uwe Hasenbein and Prof. C.-W. Wallesch (Institute of Neurological and Neurosurgical Rehabilitation, Magdeburg, Germany) for providing us with the neurological background and for letting us use their data.

References

Anderson, J.A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society, Series B*, **46**, 1–30.

Davidian, M. and Giltinan, D.M. (1995). *Nonlinear Models for Repeated Measurement Data*. London: Chapman & Hall.

Hartzel, J., Agresti, A., and Caffo, B. (2001). Multinomial logit random effects models. *Statistical Modelling*, **1**, 81–102.

Hasenbein, U., Kuss, O., Bäumer, M., Schert, C., Schneider, H., and Wallesch, C.W. (2003). Physicians' preferences and expectations in traumatic brain injury rehabilitation - results of a case-based questionnaire survey. *Disability and Rehabilitation*, **25**, 136–142.

Hedeker, D. and Gibbons, R.D. (1994). A random effects ordinal regression model for multilevel analysis. *Biometrics*, **50**, 933–944.

Tutz, G. and Hennevogl, W. (1996). Random effects in ordinal regression models. *Computational Statistics and Data Analysis*, **22**, 537–557.

A Multiple Imputation Approach to Estimation in a Gamma Frailty Model with Clustered Interval-Censored Data

K. F. Lam[1] and Tak Lun Cheung[1]

[1] Department of Statistics & Actuarial Science, The University of Hong Kong, Hong Kong, China

Abstract: The estimation of the Gamma frailty model with clustered interval censored data is considered in this study. A simple multiple imputation approach is proposed to estimate the regression parameter of the semiparametric Cox proportional hazards model for interval censored data in the univariate case. The basic idea is to iterate between the following two steps. With an additional parametric assumption on the baseline hazard function, we first impute an exact failure time to each finite interval-censored time using the approximate conditional posterior distribution. Secondly, the standard Cox partial likelihood is applied to the imputed data and the estimate of the regression parameter is updated. A robust variance estimator is also provided. Empirical results show that the proposed method works extremely well. To accommodate clustered interval censored data, an extension of the proposed method to estimate the regression and dependence parameters of a Gamma frailty model is studied. In this part, the EM algorithm for the Gamma frailty model suggested by Klein (1992, Biometrics 48, 795-806) is used instead of the partial likelihood in the aforementioned second step. The proposed method is applied to several real life data sets and the performance of the proposed method is studied via simulation. As a side product, we also propose a robust estimation procedure using a marginal approach, to make inference on the regression parameter and to compare with that based on the Gamma frailty model.

Keywords: Gamma frailty model; Clustered interval censored; Multiple imputation; Proportional hazards model.

To model interval-censored data by Cox's semiparametric proportional hazards model, we propose a simple multiple imputation approach to estimate the regression parameter in the absence of the rankings of the failure times. The basic idea is to iterate between the following two steps. With an additional Weibull assumption on the baseline hazard function, we first impute an exact failure time to each finite interval-censored time using the approximate conditional posterior distribution. Secondly, the standard partial

likelihood is applied to the imputed data and the estimate of the regression parameter is updated. The two steps are performed iteratively until convergence is achieved. Robust variance estimator for the regression parameter is also suggested to address for the misspecification of the baseline hazard function. Simulation studies showed that the proposed method performs extremely well even when the baseline hazard function is piecewise constant (See Table 1).

A study is carried out to compare the cosmetic effects of radiotherapy alone ($X = 0$) versus radiotherapy and adjuvant chemotherapy ($X = 1$) on women with early breast cancer. The variable of interest is the time to cosmetic deterioration of the patients. To compare the two treatment regimes, 46 radiation only and 48 radiation plus chemotherapy patients are considered. Patients are under intense observation in the initial 4 to 6 months after the treatments, but, when they begin to recover, the interval between visits is lengthened. Due to the fact that patients are examined only at these random times, we do not know the exact time of breast retraction, but is known to fall within the interval between two consecutive visits. The estimated regression parameter using our method together with the estimates by other existing methods (reproduced from Pan 2000 and Betensky et al. 2002) are tabulated in Table 2. Our estimates are very similar to the others.

Correlated survival data are often observed when failure times are collected on clusters of items or individuals. To accommodate clustered interval-censored data, an extension of the proposed method to estimate the regression and dependence parameters of a Gamma frailty model is studied. In this part, we impute the survival times using the approximate joint conditional posterior distribution, and the EM algorithm for the Gamma frailty model suggested by Klein (1992) is used instead of the partial likelihood in the second step of the univariate setup. The performance of the proposed method is studied via simulation (See Table 3). Again, the proposed methodology works extremely well except for a not so alarming underestimation in the dependence parameter, θ.

For illustration, we apply our proposed method to analyze the diabetic retinopathy study (DRS) data. The main purpose of the study is to assess the effectiveness of laser photocoagulation in delaying the time to onset of blindness in patients with diabetic retinopathy. It is also of interest to examine whether the effect, if it exists, depends on the type of diabetes, namely juvenile versus adult diabetes. One eye of each patient is randomly selected for treatment and the other eye is treated as control. As the failure times for both eyes are correlated, multivariate survival analysis is desired. The endpoint used to assess the treatment effect is the occurrence of visual acuity less than 5/200 at two consecutively completed 4-month follow-ups. Hence, the occurrence times are interval-censored. We analyze the data by using the proposed method. As a side product, we also propose a robust estimation procedure using a marginal approach to make inference on the

regression parameter and to compare with that based on the Gamma frailty model. The results are tabulated in Table 4.

A simple multiple imputation approach is proposed to model univariate interval-censored data. The main advantage of this method over the others is that it can be extended easily to accommodate clustered interval-censored data as illustrated in the example.

TABLE 1. *Simulation Results (Univariate Case) with imputation size = 10, 400 replications and true* $\beta = 1.000$ *(Ct=constant).*

Robust variance estimator?	True Baseline	$\hat{\beta}$	Empirical SD($\hat{\beta}$)	Average SE($\hat{\beta}$)	Coverage (95% C.I.)
No	Weibull	1.043	0.268	0.273	95.75%
	Piecewise Ct	0.993	0.261	0.271	95.75%
Yes	Weibull	1.019	0.278	0.269	95.50%
	Piecewise Ct	0.998	0.272	0.267	95.00%

TABLE 2. *Treatment effect for breast cosmesis data using proportional hazards model with various methods.*

Model	Estimate	Standard Error
Exponential	0.742	0.277
Mid-point Imputation	0.839	0.286
Huang and Wellner (1995)	0.795	0.29
Finkelstein (1986)	0.791	0.288
Satten (1996)	0.890	0.297
Satten et al. (1998)	0.878	0.294
Goggins et al. (1998)	1.450	0.371
Pan (2000)	0.90	0.29
Betensky et al. (2002)	1.053	0.270
Our results	0.849	0.287

TABLE 3. *Simulation Results (Multivariate Case) with imputation size = 500 and 500 replications (TCB=True Conditional Baseline).*

TCB	α	True	$\hat{\alpha}$	Empirical SD($\hat{\alpha}$)	Average SE($\hat{\alpha}$)	Coverage (95% C.I.)
Weibull	β	-0.693	-0.701	0.218	0.204	93.2%
	θ	0.500	0.406	0.225	0.244	90.2%
	β	-0.693	-0.675	0.244	0.210	90.8%
	θ	1.000	0.871	0.317	0.318	89.0%
	β	-0.693	-0.645	0.243	0.202	88.8%
	θ	2.000	1.851	0.463	0.447	90.6%
Piecewise constant hazards	β	-0.693	-0.681	0.234	0.208	92.0%
	θ	0.500	0.386	0.224	0.256	90.0%
	β	-0.693	-0.675	0.234	0.208	92.4%
	θ	1.000	0.846	0.303	0.320	89.6%
	β	-0.693	-0.653	0.252	0.205	89.2%
	θ	2.000	1.752	0.434	0.440	88.4%

TABLE 4. *Parameters estimates (with robust standard error between brackets) of diabetic retinopathy study using various methods and models. MA=marginal approach. FM=Frailty Model*

Method	Covariates			Frailty
	Type	Treatment	Interaction	
Huster et al. (1989)	0.37	-0.43	-0.84	2.01
	(0.20)	(0.18)	(0.30)	(0.34)
Liang et al. (1993)	0.34	-0.42	-0.84	- -
	(0.20)	(0.19)	(0.30)	
Lin (1994)	0.34	-0.43	-0.85	- -
	(0.20)	(0.19)	(0.30)	
Ross and Moore (1999)	0.35	-0.44	-0.84	2.04
	(0.21)	(0.18)	(0.29)	(0.35)
Our results (MA)	0.37	-0.41	-0.88	- -
	(0.20)	(0.22)	(0.35)	
Our results (FM)	0.43	-0.51	-0.99	2.07
	(0.22)	(0.23)	(0.40)	(0.35)

References

Betensky, R.A., Lindsey, J.C., Ryan, L.M., and Wand, M.P. (2002). A local likelihood proportional hazards model for interval censored data. *Statistics in Medicine*, **21**, 263–275.

Klein, J.P. (1992). Semi-parametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, **48**, 795–806.

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

Hierarchical Generalized Linear Models

Youngjo Lee[1]

[1] Department of Statistics, Seoul National University, Shillim-dong, Kwanak-ku, Seoul, Korea 151-742. Email: youngjo@plaza.snu.ac.kr

Abstract: Hierarchical generalized linear models (HGLMs) are developed as a synthesis of (i) generalized linear models (GLMs) (ii) mixed linear models, (iii) joint modelling of mean and dispersion and (iv) modelling of spatial and temporal correlations. Statistical inferences for complicated phenomena can be made from such a HGLM, which is capable of being decomposed into component GLMs, allowing the application of standard GLM procedures to those components, in particular those for model checking.

Keywords: Hierarchical generalized linear models; HGLMs; Hierarchical likelihood; H-likelihood.

1 Introduction

There have been many models and many methods proposed for the description and analysis of correlated non-normal data. One general approach is via the use of random effect models. In 1996 Lee and Nelder introduced a class of models called HGLMs, and recently Lee and Nelder (2001a) presented them as a synthesis of three widely-used existing model classes, (i) GLMs (McCullagh and Nelder, 1989), (ii) mixed linear models having both fixed and random effects (Longford, 1993) and (iii) models with structured dispersions as used in the analysis of data from quality improvement experiments (Nelder and Lee, 1991).
In Lee and Nelder (2001b) we introduced an extended HGLM for modelling and analysing spatial and temporal correlations for correlated non-normal data. This covers a broad class of models, and we show that many previously developed models appear as instances of our model. Rich classes of correlated patterns in non-Gaussian models can be produced via HGLMs without requiring explicit multivariate generalizations of non-Gaussian distributions. In this talk we summarize inference from HGLMs and list practical applications.

2 H-likelihood Inference

For inference in HGLMs Lee and Nelder (1996) proposed the use of an h-likelihood of the form

$$h = h(\beta, \sigma) = \log\{f(y|v;\beta)\} + \log\{f(v;\sigma)\}, \tag{1}$$

where $f(y|v;\beta)$ and $f(v;\sigma)$ denote the conditional density function of the response y given v and the density function of v, respectively, and β is a regression parameter. In forming the h-likelihood the choice of the scale of random effects is important. Note that v is the scale on which the random effects are assumed to occur linearly in the linear predictor.
By contrast, the marginal likelihood m can be obtained by integrating out the random effects from the h-likelihood:

$$m = \log\left\{\int \exp(h)\ dv\right\}. \tag{2}$$

However, integration becomes more difficult as the number of random components increases. An important advantage of the h-likelihood approach is that it facilitates inference in models with complex random effect structures *without recourse to integration.*
Let l be a likelihood, either a marginal likelihood m or an h-likelihood h, with nuisance parameters θ. Lee and Nelder (2001a) considered a function $p_\theta(l)$, defined by

$$p_\theta(l) = [l - \frac{1}{2}\log\det\{A(l,\theta)/(2\pi)\}]|_{\theta=\widehat{\theta}} \tag{3}$$

where $A(l,\theta) = -\partial^2 l/\partial\theta^2$ and $\widehat{\theta}$ solves $\partial l/\partial\theta = 0$. For fixed effect parameters β the use of $m_P \equiv p_\beta(m)$ is equivalent to conditioning on $\widehat{\beta}$ (Cox and Reid, 1987), while for random effects v the use of $p_v(h)$ is equivalent to integrating them out by using the Laplace approximation. Lee and Nelder (2001a) showed that $h_P \equiv p_\tau(h)$ where $\tau = (\beta^T, v^T)^T$ is approximately $p_\beta(p_v(h))$; in general $m \approx p_v(h)$ and $m_P \approx h_P$. In mixed linear models h_P becomes Harville's (1977) restricted likelihood m_P, and thus h_P is a natural extension of the restricted likelihood for dispersion components in linear mixed models to non-normal mixed-effect models. Therefore, h_P may be viewed as a proper likelihood for the dispersion parameters σ after eliminating the nuisance parameters τ.

3 Remarks on Inference in Hierarchical Models and H-likelihood

For inference in hierarchical models, a considerable amount of effort has been devoted to implementing methods based upon marginal likelihood,

which becomes computationally heavier as the number of random components increases. This difficulty has limited the wider application of HGLM-type models.

It is perhaps unfortunate that Bayesians, from Lindley and Smith (1972) onwards, seem to have made a major play for the high ground in all hierarchical modelling, implying, effectively, that the Bayesian approach is *the* method of choice when dealing with hierarchical models. The availability of MCMC, making many problems seem more solvable via Bayesian computations, has appeared to justify this point of view.

For inference in HGLMs, Lee and Nelder (1996) proposed the use of hierarchical likelihood (or h-likelihood) defined at (1). By using h-likelihood, we may deal with such models directly because there is an explicit analytical form for this type of likelihood (Nelder, in press). H-likelihood will, we believe, become widely used for inference in hierarchical models as it is a natural extension of Fisher likelihood to models with random parameters. Moreover, h-likelihood estimation is based on a statistically and numerically efficient fitting algorithm which provides a straightforward REML extension for inference on dispersion parameters. Finally we note that subject-specific inference is possible without resorting to an empirical Bayesian framework.

Despite these obvious strengths, one apparent criticism of the h-likelihood method derives from a belief that h-likelihood provides qualitatively different (i.e. non-invariant) inferences for trivial re-expressions of the underlying model. This perspective is due to a misunderstanding of the nature of h-likelihood: see Lee and Nelder (2003a) for detailed discussion. Another criticism of the h-likelihood method is its bias in parameter estimators for binary data. Yun and Lee (2003) showed that if it is properly implemented there is no such bias and it gives better estimators than the marginal likelihood method using Gauss-Hermite quadrature.

4 Applications of H-likelihood

The scope of the h-likelihood paradigm is wide and continues to expand - as evidenced by other contributions to this workshop. Already, the use of h-likelihood provides new solutions to various problems. These include:

1. Joint modelling of mean and dispersion (Lee and Nelder, 2001a),
2. The analysis of temporally and spatially correlated data (Lee and Nelder, 2001b),
3. A new class of models for stochastic volatility in the finance area (Lee and Nelder 2003b),
4. The provision of model checking to see if the postulated pattern of random effects is supported by the data (Lee and Nelder, 2001a)

5. Meta analysis (Lee and Nelder, 2002),
6. Analysis of survival data (Ha, Lee and Song, 2001; Ha, Lee and Song, 2002),
7. Implicit implementation of an EM-type algorithm to yield good estimators for censored linear mixed models (Ha and Lee, 2003),
8. The prediction of future observations (Pawitan, 2001, Chapter 16),
9. A simple alternative to kernel smoothing (Pawitan, 2001),
10. A new way of modelling long-range dependence and self-similarity processes for internet queuing systems (Sohn, Yun and Lee (2003),
11. New robust sandwich variance estimates for fixed effect estimators, which cannot be obtained from marginal likelihood (Lee, 2002), and
12. Alternatives to generalized estimating equations, based on extended likelihood rather than the *ad hoc* approach of generalized estimating equations (Zeger et al, 1988).

5 Discussion

The concept and definition of h-likelihood is reviewed and its several virtues extolled. It is a major competitor for existing marginal (and usually Bayesian) methods for inference in the hierarchical modelling paradigm and its use is destined to become routine and to displace existing MCMC methods in a large family of relevant statistical models.

References

Ha, I.D., Lee, Y., and Song, J.K. (2001). Hierarchical likelihood approach for frailty models. *Biometrika*, **88**, 233–243.

Ha, I.D., Lee, Y., and Song, J.K. (2002). Hierarchical likelihood approach for mixed linear models with censored data. *Lifetime Data Analysis*, **8**, 163–176.

Ha, I.D. and Lee, Y. (2003). Multilevel mixed linear models for survival data. Submitted.

Lee, Y. (2002) Robust variance estimators for fixed-effect estimates with hierarchical likelihood. *Statistics and Computing*, **12**, 201–207.

Lee Y. and Nelder J.A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B*, **58**, 619–678.

Lee, Y. and Nelder, J.A. (2001a). HGLMs: a synthesis of GLMs, random effect models and structured dispersions. *Biometrika*, **88**, 987–1006.

Lee, Y. and Nelder, J.A. (2001b). Modelling and analysing correlated non-normal data. *Statistical Modelling*, **1**, 7-16.

Lee, Y. and Nelder, J.A. (2002). Analysis of the ulcer data using hierarchical generalized linear models. *Statistics in Medicine*, **21**, 191–202.

Lee, Y. and Nelder, J.A. (2003a). Hierarchical likelihood and invariance. Submitted.

Lee, Y. and Nelder, J.A. (2003b). Double hierarchical generalized linear models. Submitted.

Lindley, D.V. and Smith, A.F.M. (1972). Bayesian estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 1–41.

Nelder, J.A. (2003). Extended likelihood inference applied to a new class of 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. 335–338.

Pawitan, Y. (2001). *In all Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Clarendon Press.

Sohn, S.Y., Yun, S., and Lee Y. (2003). Modeling a non-homogeneous LRD queuing system with covariates: Inverse gamma mixture of Pareto. Submitted.

Yun, S. and Lee, Y. (2003). Comparison of hierarchical and marginal likelihood estimators for binary outcomes. *Computational Statistics and Data Analysis*. (in press).

Zeger, S.L., Liang, K.Y., and Albert, P.S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, **44**, 1049–1060.

The Additive Genetic Gamma Frailty Models for Genetic Linkage and Association Analysis

Hongzhe Li[1] and Xiaoyun Zhong[1]

[1] University of California, Davis, California, 95616, USA

Abstract: We introduce and demonstrate the application of the additive genetic gamma frailty models for genetic linkage and association analysis. Such models are developed in order to account for disease phenotypic or etiological heterogeneity, including variable age of onset and possible environmental risk factors. Both real data sets and simulations indicate that the methods can potentially gain power in mapping genes for complex human diseases.

Keywords: Linkage analysis; Association analysis; Age of onset; Frailty model.

1 Introduction

Many complex human diseases are due to multiple disease genes and both genetic and environmental risk factors. These diseases often also show variable age of disease onset. Examples include human cancers such as breast cancer and prostate cancer. Early age of cancer onset is a strong indicator for genetic predisposition and variable age of onset is often a good indicator for disease heterogeneity. Therefore, for complex diseases with variable age of onset, it is important to incorporate these information into genetic linkage and association analysis.

In order to incorporate both covariates and age of onset information into genetic analysis, we have defined an additive genetic gamma frailty model constructed based on the inheritance vectors (Li and Zhong, 2002). Unlike the previously proposed frailty models, our models construct the frailties based on gene segregation within a family. From the conditional frailty model, we further derived the joint survival functions for age of onset data within a family, and explicitly obtained the conditional hazard ratio for sib pairs who share different number of allele identify by descent at the putative disease locus. Within this modelling framework, we derive a retrospective likelihood ratio test for linkage and a score test for genetic association in the linked region using sibships data.

The paper is organized as follows: we first define the model and formulate tests of linkage and association in terms of the parameters in this model. We then present results of analysis of the data sets from the 12th Genetic Analysis Workshop. We conclude the paper with a brief discussion.

2 Statistical Models

2.1 The Additive Genetic Gamma Frailty Model for Age of Onset

Consider a sibship with n sibs. Let T_j be the random variable of age at disease onset for the jth sib. Let (t_j, δ_j) be the observed data where t_j is the observed age at onset if $\delta_j = 1$, and age at censoring if $\delta_j = 0$. Consider a candidate marker d in the linked region, and let $g = (g_1, \cdots, g_n)$ denote the vector of genotypes at locus d of the m family members of known age at disease onset. We assume that the hazard function of developing disease for the jth individual at age t_j is modelled by the proportional hazards model with random effect Z_j,

$$\lambda_j(t_j|Z_j) = \lambda_0(t_j)\exp(X_{g_j}\beta)Z_j, \text{ for } j = 1, 2, \cdots, n, \tag{1}$$

where $\lambda_0(t)$ is the unspecified baseline hazard function, and X_{g_j} denotes some function of the jth offspring's marker genotype in the family, for example, for additive model, $X_{g_j} = l$, $l = 0, 1, 2$, counts the number of the putative high-risk allele D and is for the genotype of jth member in the family who carries l copies of the putative high-risk allele D. Z_j is the unobserved genetic frailty. Following Li and Zhong (2002), we define the genetic frailty as the following

$$Z_j \quad = \quad U_{dv_{2j-1}} + U_{dv_{2j}} + U_p,$$

where $V_d = (v_1, v_2, \cdots, v_{2n-1}, v_{2n})$ is the inheritance vectors (Lander and Green, 1987) of a sibship at d locus, $v_{2j-1} = 1$ or 2, and $v_{2j} = 3$ or 4 for $j = 1, 2, \cdots, n$. The inheritance vector indicates which parts of the genome at locus d are transmitted to the n children from the father and the mother. Here U_{d1} and U_{d2} are used to represent the genetic frailties due to part of the genome on the two chromosomes of the father at locus d, and U_{d3}, and U_{d4} are analogous though for the mother. The random frailty term, U_p, takes into account possible genetic contributions to the disease due to loci unlinked to locus d, or contributions to shared familial effects. Assume that the U_{d1}, U_{d2}, U_{d3} and U_{d4} are independently and identically distributed across different families as $\Gamma(\nu_d/2, \eta)$, and U_p is distributed as $\Gamma(\nu_p, \eta)$ over different sibships, where η is the inverse scale parameter and ν_d and ν_p are the shape parameters. To make the baseline hazard $\lambda_0(t)$ identifiable, let $\nu_d + \nu_p = \eta$. Under this restriction, there are two free parameters, ν_d and ν_p, and $U_{di} \sim \Gamma(\nu_d/2, \nu_d + \nu_p)$, $U_p \sim \Gamma(\nu_p, \nu_d + \nu_p)$, $i = 1, \cdots, 2n$ and $Z_j \sim \Gamma(\nu_d + \nu_p, \nu_d + \nu_p)$.

Li and Zhong (2002) considered a similar model as (1), but they did not include the $X_{g_j}\beta$ term in the model. They further showed that the null hypothesis that the candidate locus does not contribute to the risk of disease can be formulated as testing $H_0 : \nu_d = 0$. Li and Zhong (2002) gave

a retrospective likelihood ratio based test assuming that the population disease rates are known. Li (2002) gave a prospective likelihood ratio based test using the EM algorithm.

2.2 A score Test for Genetic Association in the Linked Region

Once linkage has been established, more markers are usually typed in this region and genetic association test based on linkage disequilibrium is often performed. However, it is well known that the transmission of alleles to different sibs within a family is dependent in the linked region (Ewens and Spielman, 1995). Therefore, direct applications of some of the tests treating sibs within a family as independent will result in inflated type 1 error rates. As we can see, when $\beta = 0$, the hazard function (1) and the joint density and survival function for a sibship does not depend on the genotype at the locus d, therefore, test of allelic association between locus d and the disease or the null hypothesis that genotype at candidate locus will not affect the risk of the disease can be formulated as testing $H_0 : \beta = 0$.
Let $M_i = (g_{i1}, \cdots, g_{in_i})$ be the vector of the marker genotypes at the candidate marker locus for the n_i children in the ith family, and $g_i = (g_{iF}, g_{iM})$ be the vector of parental marker genotypes. We can derive a score test based on the following retrospective conditional likelihood for the ith family,

$$L_i(\nu_d, \nu_p, \Lambda_0(t), \beta; g_i) = Pr(M_i|t_i, \delta_i, g_i) = \frac{Pr(M_i|g_i)Pr(t_i, \delta_i|M_i)}{\sum_M Pr(M|g_i)Pr(t_i, \delta_i|M)},$$

where $\sum_M$ denotes summation over all possible offspring genotype vectors M. The corresponding score statistic can be written as

$$S_i = \sum_{j=1}^{n_i} [\delta_{i_j} - \Lambda_0(t_{i_j})F_{i_j}(t_i, \delta_i, \nu)](X_{g_{i_j}} - E(X_{g_{i_j}}|g_{iF}, g_{iM})),$$

where $F_{i_j}(t_i, \delta_i, \nu)$ is a function of the age of onset data and the linkage parameters. The score test can then be defined as $T = \sum_{i=1}^{N} S_i / \sqrt{\sum_{i=1}^{N} V(S_i)}$.

3 Applications to the 12th Genetic Analysis Workshop Data

To demonstrate the proposed methods, we analyzed the simulated data from the general population of 12th Genetic Analysis Workshop (GAW12). The true disease model includes seven major genes which influence the disease liability and age of onset. Among these 7 genes, only major gene 7 directly contributes to age of onset and major gene 6 directly contributes

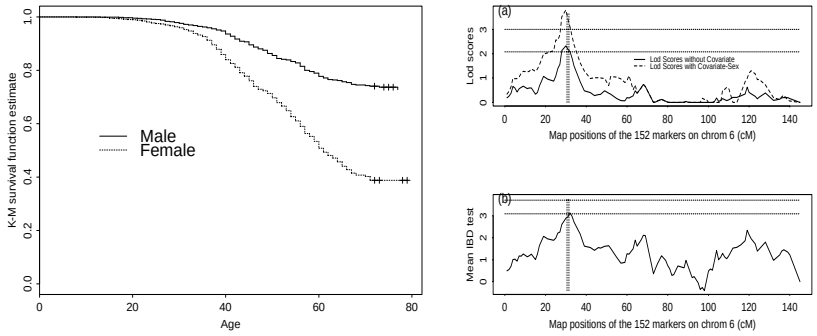


FIGURE 1. *Left plot: Kaplan-Meier disease-free survival curves for males and females estimated based on the founders' data; right plot: Linkage results based on our method (top plot) and the mean IBD test for chromosome 6 (bottom plot), where the dashed horizontal lines refer to the critical values corresponding to 0.001 and 0.0001 significant levels respectively and the dashed vertical lines mark the location of major genes 6 and 7.*

to disease liability. Both major gene 6 and 7 reside on Chromosome 6, with major gene 6 on the 30.5cM position and major gene 7 on the 31.5cM position. Our analysis focuses on chromosome 6, which incudes a total of 152 microsatellite markers of an average of roughly 1 cM apart. There is a total of 50 replicates, each containing 23 extended pedigrees with 1,497 total individuals. We used the first thirty replicates of simulated data sets from the general population, extracting 500 affected sib pairs with their parents from each pedigree to ensure independence between nuclear families. We then calculated the Kaplan-Meier nonparametric survival estimate as the approximation of the baseline hazard function using the available age of onset data from all the founders of the first thirty replicates (see left plot of Figure 1). The plot indicates difference in survival rates between males and females.

We first performed linkage analysis for chromosome 6. The right panel of Figure 1 shows the results. These plots indicate that our methods give stronger evidence of linkage than the mean IBD test, and adjusting for sex as a covariate improves the power. For the regions far away from the true disease region, no significant linkage signal is observed.

We then performed genetic association analysis for all the 65 SNPs existed in the coding region of major gene 6. Figure 2 plots the negative logarithms of the p-values of the test versus the sequence number of the 65 SNPs. We observed that some SNPs in the coding region showed a significant evidence of association with the disease, especially when the sex covariate was taken

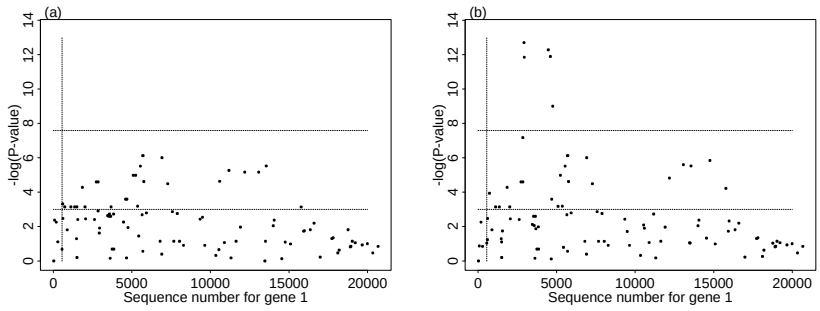


FIGURE 2. *Results of association tests for SNPs in gene 1, (a) without adjusting for sex; (b) adjusting for sex. The horizontal lines correspond to the significance levels of 0.001 and 0.0001 and the vertical line indicate the true disease variant.*

into account in the analysis. Note that we did not observe strong association for some of the SNPs which are close to the true disease variant. This can happen since the linkage disequilibrium is not just a function of the distance between the two locus, it also depends on the allele frequencies.

4 Conclusions and Discussions

In conclusion, the statistical tests based on the additive genetic gamma frailty models provide a flexible framework for incorporating age of onset and environmental risk factors into genetic analysis of complex diseases. The proposed methods are allele-sharing based and do not require specification of the mode of inheritance or the penetrance functions. Analysis of the GAW12 data sets and our simulation studies indicate that the methods are applicable to real data sets (Zhong and Li, 2003, in preparation).

References

Ewen, W.J. and Spielman, R.S. (1995). The transmission/disequilibrium test: History, subdivision and admixture. *American Journal of Human Genetics*, **57**, 455–464.

Lander, E. and Green, P. (1987). Construction of multilocus genetic maps in humans. *Proceedings of National Academy of Sciences USA*, **84**, 2363–2367.

Li, H. (2002). The additive genetic gamma frailty model for linkage analysis of diseases with variable age of onset using nuclear families. *Lifetime Data Analysis*, **8**, 315–334.

Li, H. and Zhong, X. (2002). Multivariate survival models induced by genetic frailties, with application to linkage analysis. *Biostatistics*, **3**(1), 57–75.

Multilevel Structural Equation Models: The Limited Information and the Multivariate Multilevel Approach

Cora Maas[1] and Joop Hox[1]

[1] Department of Methodology and Statistics, Utrecht University, PO Box 80140, 3508 TC Utrecht, The Netherlands, C.Maas@fss.uu.nl

Abstract: *Structural equation modeling*, or *SEM*, is a general and convenient framework for statistical analysis that includes as special cases several traditional multivariate procedures, such as factor analysis, multiple regression analysis, discriminant analysis, and canonical correlation. Structural equation models for multilevel data have been formulated by several authors. The approach to multilevel SEM outlined by Muthn is particularly interesting, because he shows that structural equation modeling of multilevel data is possible using available standard SEM software. A different approach is to estimate the covariance matrices at the distinct levels, using standard multilevel regression software, as proposed by Goldstein. This approach has the advantage that it also uses standard SEM software, but the models and hence the program setups are far less complicated than the models and setups implied by the Muthn approach. This paper examines both approaches in some detail, and compares them on an exemplary data set.

Keywords: Multilevel structural equation models.

1 The Muthén Approach: Decomposing Multilevel Variables

Multilevel structural models assume that we have a population of individuals that are divided into groups. The individual data are collected in a p-variate vector Y_{ig} (subscript i for individuals, g for groups). The variates Y_{ig} can be decomposed into a between groups component $Y_B = \overline{Y}_g$, and a within groups component $Y_W = Y_{ig} - \overline{Y}_g$. This decomposition leads to a between groups covariance matrix Σ_B (the population covariance matrix of the disaggregated group means Y_B) and a within groups covariance matrix Σ_W (the population covariance matrix of the individual deviations from the group means Y_W). Following the same logic, we can also decompose the sample data, which leads to the sample covariance matrices S_B and S_W. An unbiased estimate of the population within groups covariance matrix Σ_W is given by the pooled within groups covariance matrix S_{PW}. For computational reasons it is convenient to calculate not the between groups

covariance matrix S_B itself but the *scaled between groups covariance matrix* for the disaggregated group means S_B^*, which in the balanced case equals nS_B. Muthén (1989, 1990) shows that S_{PW} is the maximum likelihood estimator of Σ_W, with sample size $N-G$, and S_B^* is the maximum likelihood estimator of the composite $\Sigma_W + c\Sigma_B$.
In Muthén's approach, we use the multi-group option of conventional SEM software for a simultaneous analysis at both levels. We specify two groups, with covariance matrices S_{PW} and S_B^*. The model for Σ_W must be specified for both S_{PW} and S_B^*, with equality restrictions between both 'groups' to guarantee that we are indeed estimating the same model in both covariance matrices, and the model for Σ_B is specified for S_B^* only, with the scale factor $c = n$ built into the model.

2 The Multivariate Multilevel Approach: Direct Estimation of the Covariance Matrix at each Level

Goldstein (1987, 1995) suggests using a multivariate multilevel (MVML) regression model to produce a covariance matrix at the different levels, and to input these in a second step into a standard SEM program for further analysis. Multivariate multilevel regression models are multilevel regression models that contain more than one response variable.
In multivariate multilevel models, the variables constitute the lowest-level units. In most applications, the variables would be the first level, the individuals the second level, and if there are groups, these form the third level. If we have p response variables, Y_{hij} is the response on measure h of individual i in group j. We define p dummy variables, one for each response variable. In the multivariate multilevel model, the fixed part contains p regression coefficients for the dummy variables, which are the p overall means for the p outcome variables. The random part contains two covariance matrices, Σ_{ij} and Σ_j, which contain the variances and the covariances of the regression slopes for the dummies on the individual and the group level. Since that individual level and group level covariances are estimated directly, they can be modeled directly and separately by any SEM program. As a result, we get separate model tests and fit indices at all levels. The multivariate multilevel approach to multilevel SEM also generalizes straightforwardly to more than two levels. The resulting simplicity is a distinct advantage of the multivariate multilevel approach. There are other advantages as well. First, since the multilevel multivariate model does not assume that we have a complete set of variables for each individual, incomplete data are accommodated without special effort. Second, if we have dichotomous variables, we can use the multilevel generalized linear model to produce the covariance matrices, again without special effort.

TABLE 1. *Population and estimated values of the model parameters*

Loadings	Population			Muthén method			MVML method		
Variables	F1	F2	BF1	WF1	WF2	BF1	WF1	WF2	BF1
X1	.3		.5	.300		.497	.300		.499
X2	.4		.4	.400		.395	.400		.397
X3	.5		.3	.500		.293	.500		.295
X4		.3	.5		.300	.497		.300	.499
X5		.4	.4		.400	.395		.400	.397
X6		.5	.3		.500	.293		.500	.295

3 Comparing the Two Approaches

In this paper, we compare the limited information Muthén approach and the multivariate multilevel approach to multilevel SEM. Our benchmark model is a two-level factor model with six observed variables, one factor at the group level, and two factors at the individual level. Using procedures outlined by Waller we have constructed a two-level data set that *exactly* reproduces the benchmark model. The multilevel structure has 100 groups all of size 50. The group size and the number of groups have been chosen to be both large enough to ensure accurate estimation of both parameters and standard errors at all levels (cf. Hox & Maas, 2001).
Since the data are balanced, Muthén's method in this case is a full information Maximum Likelihood method. The multivariate multilevel approach produces Maximum Likelihood estimates, and since the input are two covariance matrices estimated by Maximum Likelihood methods the results should be comparable to the estimates produced by the Muthén method.

4 Results

Table 1 shows the population values of the factor loadings and variances in the two-level factor model. In addition, it shows the estimates produced by Muthén's method and by the multivariate multilevel (MVML) method. It is clear from Table 1 that the individual level loadings are estimated with total accuracy. Since the individual-level sample size is $N-G = 4900$, this is not surprising. At the group level, where the sample size is 100, the loading estimates are very close to the population values. The estimates produced by the MVML method are somewhat closer to the known population values, but both sets of estimates are so close to the true values that this difference is utterly trivial.
Table 2 shows the population values of the factor loadings and the standard errors produced by Muthén's method and by the multivariate multilevel (MVML) method.

TABLE 2. *Population and estimated standard errors of the model parameters*

Loadings	Population			Muthén method			MVML method		
Variables	F1	F2	BF1	WF1	WF2	BF1	WF1	WF2	BF1
X1	.3		.5	.010		.006	.010		.101
X2	.4		.4	.012		.008	.012		.095
X3	.5		.3	.014		.092	.014		.090
X4		.3	.5		.010	.006		.010	.101
X5		.4	.4		.012	.008		.012	.095
X6		.5	.3		.014	.092		.014	.090

It is clear from Table 2 that the standard errors of the individual level loadings are estimated with total accuracy. Since the individual-level sample size is $N - G = 4900$, this is not surprising. At the group level, where the sample size is 100, the standard errors are produced by the Muthén and MVML method are a bit different. In this case, they would both lead to the same conclusion, but the p-values and confidence intervals are certainly not the same. To check the standard errors, we carried out a parametric bootstrap on the covariance matrices produced by the MVML method. The bootstrapped standard errors are very close to the asymptotic standard errors produced by the direct estimation using the MVML method.

References

Goldstein, H. (1987). *Multilevel Models in Educational and Social Research*. London: Griffin.

Goldstein, H. (1995). *Multilevel Statistical Models*. London: Edward Arnold.

Hox, J.J. and Maas, C.J.M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling*, **8**, 157–174.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, **54**, 557–585.

Muthén, B. (1990). *Means and Covariance Structure Analysis of Hierarchical Data*. Los Angeles: UCLA Statistics series, **62**.

Non-PH Multivariate Survival Models Based on the GTDL

Gilbert MacKenzie[1], Il Do Ha[2], and Youngjo Lee[3]

[1] Centre for Medical Statistics, Keele University, Keele, Staffordshire ST5 5BG, UK. Email: g.mackenzie@keele.ac.uk
[2] Faculty of Information Science, Kyungsan University, Kyungsan, 712-240, South Korea. Email: idha@kyungsan.ac.kr
[3] Department of Statistics, Seoul National University, Seoul, 151-742, South Korea. Email: youngjo@plaza.snu.ac.kr

Abstract: Correlated survival times may be modelled by introducing a random effect, or frailty, component into the hazard function. For multivariate survival data we extend a non-PH model, the generalized time-dependent logistic (GTDL) survival model (MacKenzie, 1996, 1997), to include random effects. The extension leads to two different, but related, non-PH models according to the method of incorporating the random effects. The h-likelihood procedures of Ha, Lee and Song (2001) and Ha and Lee (2003), which obviate the need for marginalization (over the random effect distribution), are derived for these extended models and their properties discussed. The new models are used to analyze two practical examples in the survival literature and the results are compared with those obtained from fitting the PH and PH frailty models.

Keywords: Frailty models; Generalized time-dependent logistic; Hierarchical-likelihood; Non-PH model; Random effect.

1 Introduction

Proportion hazards (PH) frailty models which extend the standard PH model (Cox 1972) to allow frailty are frequently to analyze multivariate survival data which may arise, for example, when recurrent or multiple event times on the same subject. However, the assumption of proportionality can sometimes be inappropriate.

In this paper we introduce a flexible non-PH random-effect model based on the generalized time-dependent logistic (GTDL) survival model (MacKenzie, 1996). The GTDL generalizes the relative risk (RR) in Cox's PH model to time-dependent form. The model, a wholly parametric competitor for the PH model, has several interesting properties including a frailty interpretation. In particular, by retaining Cox's constant of proportionality as the leading term in the RR, the model is not only capable of representing data which conform the PH assumption, but can also accommodate a

wider class of survival data in which the assumption of proportionality is untenable.
We, extend the GTDL to the multivariate survival data setting in two ways, adopt the hierarchical likelihood (h-likelihood) approach of Ha, Lee and Song (2001) and Ha and Lee (2003) for inference, use the new models to analyze two well known practical data sets from the literature and compare the results with the PH and PH frailty models.

2 The GTDL Model

A non-PH model, the GTDL regression model (MacKenzie, 1996), is defined by the hazard function:

$$\lambda(t;x) = \lambda_0 p(t;x), \tag{1}$$

where $\lambda_0 > 0$ is a scalar, $p(t;x) = \exp(t\alpha + x^T\beta)/\{1 + \exp(t\alpha + x^T\beta)\}$ is a linear logistic function in time, α is a scalar measuring the effect of time and β is a $p \times 1$ vector of regression parameters associated with fixed covariates $x = (x_1, \ldots, x_p)^T$. The relative risk (RR), the ratio of hazard rates for two subjects with different values of covariates, $x^{(1)}$ and $x^{(2)}$, is given by

$$\gamma(t; x^{(1)}, x^{(2)}) = \lambda(t; x^{(1)})/\lambda(t; x^{(2)}) = \exp\{(x^{(1)} - x^{(2)})^T\beta\}\psi(t, x^{(1)}, x^{(2)}), \tag{2}$$

where $\psi(t, x^{(1)}, x^{(2)}) = \{1+\exp(t\alpha+x^{(2)T}\beta)\}/\{1+\exp(t\alpha+x^{(1)T}\beta)\}$. The leading term on the right hand side of (2), Cox's constant RR over time, is thus moderated by $\psi(\cdot)$, a function of both time and covariates. That is, the model (1) is a non-PH, but when $\alpha = 0$ resulting RR is time invariant and model is then PH. The cumulative hazard function is given by

$$\Lambda(t;x) = \int_0^t \lambda(s;x)\ ds = \frac{\lambda_0}{\alpha}\log\left\{\frac{1+\exp(t\alpha + x^T\beta)}{1+\exp(x^T\beta)}\right\}. \tag{3}$$

Under non-informative censoring the ordinary censored-data likelihood, which depends on (1) and (3), is easily constructed.

3 Extended GTDL Models

The multivariate data structures are as follows. Let T_{ij} $(i = 1, \ldots, q,\ j = 1, \ldots, n_i,\ n = \sum_i n_i)$ be the survival time for jth observation of the ith subject. Denote by U_i the unobserved frailty (or random effect) for the ith subject.
We extend the model (1) to include a frailty term acting multiplicatively on the individual hazard rate. Given $U_i = u_i$, the conditional hazard function of T_{ij} takes the form

$$\lambda_{1ij}(t|u_i) = \lambda_{ij}(t)u_i, \tag{4}$$

The frailties U_i are assumed to be independent and identically distributed random variables with a density function depending on the frailty parameter θ, say $g(.|\theta)$. Alternatively, we may consider another natural extension of model (1), by including a random component in the linear predictor, $t\alpha + x^T\beta$, of (1). Given $U_i = u_i$, the conditional hazard function of T_{ij} is then of the form

$$\lambda_{2ij}(t|u_i) = \lambda_0 \frac{\exp(t_{ij}\alpha + x_{ij}^T\beta + u_i)}{1 + \exp(t_{ij}\alpha + x_{ij}^T\beta + u_i)} \ . \tag{5}$$

where the U_i have been defined above.
Models (4) and (5) are similar, but (4) assumes that the random effects act multiplicatively on the hazard function while (5) assumes they are additive on a generalized $\log_e$-odds scale, which is the usual $\log_e$-odds scale when $\lambda_o = 1$. While (4) is a conventional frailty model, (5) is not, although it is nevertheless of interest, since then the random effects and the fixed effects act linearly on the same scale.

The choice of $g(.|\theta)$ may be important. For h-likelihood inference, the choice of parametric form is wide (and testable), since marginalization is not required. In this paper we shall adopt the log-Normal distribution for h-likelihood inference - a choice to which inference on β is robust (Ha et al., 2001; Ha and Lee, 2003). Perhaps a more natural choice for Model (4) is the Gamma distribution, see Blagojevic, MacKenzie & Ha (2003) for a marginal approach. Alternatively, we may adopt a non-parametric mixture model.

4 H-Likelihood Estimation and Inference

Let the observable random variables be $Y_{ij} = \min(T_{ij}, C_{ij})$ and $\delta_{ij} = I(T_{ij} \leq C_{ij})$, where C_{ij} is the censoring time corresponding to T_{ij} and $I(\cdot)$ is the indicator function.
Following Ha, Lee and Song (2001), the h-likelihood for the model (4), denoted by h, is defined by

$$h = h(\alpha, \beta, \theta) = \sum_{ij} \ell_{1ij} + \sum_{i} \ell_{2i}, \tag{6}$$

where

$$\ell_{1ij} = \ell_{1ij}(\alpha, \beta; y_{ij}, \delta_{ij}|u_i) = \delta_{ij} \log \lambda_1(y_{ij}|u_i) - \Lambda_1(y_{ij}|u_i)$$

is the logarithm of the conditional density function for Y_{ij} and δ_{ij} given $U_i = u_i$, and $\ell_{2i} = \ell_{2i}(\theta; v_i)$ is the logarithm of the density function for $V_i = v(U_i) = \log(U_i)$ with parameter θ. Here v is scale on which the random effects influence the linear predictor and $v_i = v(u_i) = \log u_i$: see also Lee

and Nelder (1996). The maximum h-likelihood (MHL) estimating equations of $\tau = (\alpha, \beta^T, v^T)^T$ with $v = (v_1, \ldots, v_q)^T$ are given by

$$\partial h/\partial \tau = 0. \tag{7}$$

Note that the asymptotic covariance matrix (Ha et al., 2001) for $\widehat{\tau} - \tau$ is given by the inverse of $H = -\partial^2 h/\partial \tau^2$. For the estimation of the frailty parameter θ, we use Lee and Nelder's (1996) APHL (adjusted profile h-likelihood) h_P of θ after eliminating τ, defined by

$$h_P = h_A|_{\tau=\widehat{\tau}}\,, \tag{8}$$

where $h_A = h + \frac{1}{2}\log\{\det(2\pi H^{-1})\}$. Given estimates of τ, Lee and Nelder's (2001) REML (restricted maximum likelihood) estimating equation for θ, maximizing h_P, is given by

$$\partial h_A/\partial \alpha|_{\tau=\widehat{\tau}} = 0. \tag{9}$$

5 Results

We illustrate the use of models (4) and (5) and also their conditional forms (without frailty or random effects) and include Cox's PH and PH frailty models as comparators.

We analyze two sets of well-known multivariate survival data which have appeared in the literature. Firstly, the kidney infection data of McGilchrist and Aisbett (1991), comprising times to the first and second recurrences of infection in 38 kidney patients and consider a single fixed covariate, sex of the patients, coded 1 for female and 0 for male. Secondly, the placebo-controlled randomized trial of gamma interferon (γ–IFN) in chronic granulomatous disease (CGD) (Fleming and Harrington, 1991) in which scientific interest is focused on the effect of treatment on the (possibly multiple) recurrence times. In all, we analyze ten covariates including treatment. The results of the analyzes are shown in Tables 1 and 2 respectively (omitted). For the kidney data, the finding that femaleness is protective of recurrence is confirmed in all of the models fitted. The standard error is elevated in all frailty models suggesting that Cox Model and the non-PH models without frailty fail to account properly for the (positive) correlation between recurrence times. The α parameter in the non-PH models is not significant, suggesting that there is no serious departure from the PH assumption in these data.

The results for the CGD data are broadly similar in that the treatment effect is correctly identified by all models fitted. However, in these data, there is clear evidence of non-proportionality $\alpha \neq 0$, but the size of this effect is small. On the other hand, there is some difference in interpretation of the longitudinal covariate, which is identified by all of the non-PH models.

References

Blagojevic M., MacKenzie G., and Ha, I.D. (2003). A Comparison of non-PH & PH - Gamma frailty 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. 39–44.

Cox, D.R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.

Fleming, T.R. and Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. New York: Wiley.

Ha, I.D. and Lee, Y. (2003). Estimating frailty models via Poisson hierarchical generalized linear models. *Journal of Computational and Graphical Statistics*. (in press).

Ha, I.D., Lee, Y., and Song, J.-K. (2001). Hierarchical likelihood approach for frailty models. *Biometrika*, **88**, 233–243.

Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B*, **58**, 619–678.

Lee, Y. and Nelder, J.A. (2001). Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, **88**, 987–1006.

MacKenzie, G. (1996). Regression models for survival data: The generalised time dependent logistic family. *Journal of the Royal Statistical Society, 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.

McGilchrist, C.A. and Aisbett, C.W. (1991). Regression with frailty in survival analysis. *Biometrics*, **47**, 461–466.

Optimal Model Selection in a Joint Mean-Covariance Space

Gilbert MacKenzie[1] and Jianxin Pan[2]

[1] Centre for Medical Statistics, Keele University, Keele, Staffordshire ST5 5BG, UK. Email: g.mackenzie@keele.ac.uk
[2] Department of Mathematics, Manchester University, Manchester M13 9PL. Email: jpan@maths.man.ac.uk

Abstract: We reparametrize the marginal covariance matrix arising in longitudinal studies to model, jointly, the mean and covariance structures in terms of three polynomial function of time. We also compare model selection procedures based on regressogram estimation with those based on a direct search of the joint model space. Using a BIC-based model selection criterion to identify the optimum degree triple of the three polynomials, we show that the use of a saturated mean model is not optimal, explain why regressogram-based model estimation may mislead and give a new computational algorithm, based on a criterion involving three pairwise saturated profile likelihoods, for finding the global optimum model efficiently.

Keywords: BIC; Joint mean-covariance modelling; Profile likelihood.

1 Introduction

We model, jointly, the mean-covariance structures arising in longitudinal studies. The technique 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 autoregressive coefficients, ϕ. These unconstrained parameters are modelled, parsimoniously, as different polynomial functions of time. Pourahmadi (1999, 2000). The degrees of the polynomials adopted are suggested by means of the regressograms, derived from the sample covariance matrix, which plot the sample autoregressive coefficients and innovation variances against lag and time, respectively (Pourahmadi, 1999, 2000). We include a polynomial representation for the mean structure in order to fit a joint mean covariance model. This choice is reasonable for growth curve, longitudinal and multi-level data (Rao, 1987 and Goldstein et al, 1996). The resulting model is an augmented polynominal regression model involving three equations. Optimal

model selection then involves identifying the best integer triple representing, respectively, the degrees of the three polynomial functions for the mean structure, the autoregressive coefficients and the log innovation-variances.

2 Augmented Regression Model

Let y_{ij} be the jth of m_i measurements on the ith of n subjects and let t_{ij} be the time at which the measurement y_{ij} is made. Denote by $y_i = (y_{i1}, y_{i2}, ..., y_{im_i})'$ and $t_i = (t_{i1}, t_{i2}, ..., t_{im_i})'$ the $m_i \times 1$ vectors of responses and times of the ith subject. It is assumed that $y_i \sim N_{m_i}(\mu_i, \Sigma_i)$, where $\mu_i = (\mu_{i1}, \mu_{i2}, ..., \mu_{im_i})'$ and Σ_i are an $m_i \times 1$ vector and an $m_i \times m_i$ positive definite matrix, respectively. The mean μ_{ij} of y_{ij} can usually be modelled by a linear regression, $\mu_{ij} = x'_{ij}\beta$, where x_{ij} denotes the baseline covariates associated with the jth observation of the ith subject and β is an $(p+1) \times 1$ vector of regression coefficients. The subject-specific covariance matrix, Σ_i, may be modelled as $T_i \Sigma_i T'_i = D_i$. The below-diagonal entries of T_i are the negatives of the autoregressive coefficients, ϕ_{ijk}, in $\hat{y}_{ij} = \mu_{ij} + \sum_{k=1}^{j-1} \phi_{ijk}(y_{ik} - \mu_{ik})$, the linear least squares predictor of y_{ij} based on its predecessors $y_{i(j-1)}$, ..., y_{i1}. The diagonal entries of D_i are the innovation variances $\sigma^2_{ij} = \text{var}(y_{ij} - \hat{y}_{ij})$, where $1 \leq j \leq m_i$ and $1 \leq i \leq n$ (Pourahmadi, 1999). The parameters ϕ_{ijk} and $\varsigma_{ij} \equiv \log \sigma^2_{ij}$ are unconstrained and are modelled in an augmented regression as

$$\mu_{ij} = x'_{ij}\beta \qquad \phi_{ijk} = z'_{ijk}\gamma \qquad \varsigma_{ij} = h'_{ij}\lambda \tag{1}$$

where β, γ and λ are the parameters of interest. Then, minus twice the loglikelihood function, except for a constant, is given by

$$-2\ell = \sum_{i=1}^{n} \log |T_i^{-1} D_i T_i^{'-1}| + \sum_{i=1}^{n} r'_i T_i^{'} D_i^{-1} T_i r_i \tag{2}$$

where $r_{ij} = y_{ij} - x'_{ij}\beta$ is the jth element of $r_i = y_i - X_i\beta$, the vector of residuals, and the matrix X_i has row vectors x'_{ij} $(j = 1, 2, ..., m_i)$.
Pan & MacKenzie (2003) give an inter-dependent iteratively re-weighted least squares algorithm for computing the maximum likelihood estimates. Their algorithm is more general than Pourahmadi's procedure which is restricted to balanced longitudinal data.

3 Optimal Model Selection

For direct comparability with Pourahmadi's methods we choose, as our model selection criterion, the BIC, defined as

$$\text{BIC}(p, q, d) = -(2/n)\hat{\ell}_{\max} + (p + q + d + 3)(\log n/n) \tag{3}$$

where $\hat{\ell}_{\max} = \ell(\hat{\beta}_p, \hat{\gamma}_d, \hat{\lambda}_q)$ is the maximized loglikelihood for the models with the specified degree triple (p, q, d), and $p + q + d + 3$ is the number of parameters in the associated models, including polynomials of degree zero. The best triple, (p^*, q^*, d^*), say, satisfies

$$(p^*, q^*, d^*) = \arg \min_{(p,q,d)} \{\text{BIC}(p, q, d)\} \tag{4}$$

where p, q and d lie in the range 0 to $(m_0 - 1)$, and where $m_0 = \max_{1 \le i \le n}\{m_i\}$, when the data are unbalanced and $m_0 = m$ otherwise. We denote the corresponding value of $\hat{\ell}_{\max}$ by $\ell(\hat{\beta}_{p^*}, \hat{\gamma}_{d^*}, \hat{\lambda}_{q^*})$.

Use of (4) implies a direct search of the 3-dimensional joint model space which may be thought computationally expensive. However, a general procedure *is* required, because the simple regressogram-based model selection procedures proposed by Pourahmadi (1999) ignore the covariance structure between the parameters evident in the observed and expected information matrices (Pan & MacKenzie, 2003) and are therefore not optimal for model selection.

Accordingly, to minimize computational labour we propose an efficient search strategy to identify the global optimum model. From an appeal to profile likelihood theory (Barndorff-Nielsen, 1991), we conjecture that the optimum model may be found using three BIC-based searches involving the profile likelihoods obtained by saturating the parameter sets in pairs:

$$\begin{aligned} p_c^* &= \arg\min_p\{\text{BIC}(p, m-1, m-1)\} \\ q_c^* &= \arg\min_q\{\text{BIC}(m-1, q, m-1)\} \\ d_c^* &= \arg\min_d\{\text{BIC}(m-1, m-1, d)\} \end{aligned} \tag{5}$$

Our conjecture, tested successfully below, is that $(p^*, q^*, d^*) = (p_c^*, q_c^*, d_c^*)$. Our profile BIC algorithm reduces the number of maximizations required to find the global maximum from m^3 to $3m + 1$ in balanced longitudinal studies.

4 Example Analysis

Kenward (1987) analyzed an experiment in which cattle were assigned randomly to two treatment groups A and B, and their weights were measured 11 times over a 133 day period. Thirty animals received treatment A and another thirty received treatment B. Pourahmadi (2000) analyzed the data in group A using a saturated mean model with 11 parameters. Inspection of the sample regressograms suggested the use of two cubic polynomials for modelling the covariance structure, one for the autoregressive coefficients, in lag, and another for the innovation variances, in time. We re-analyze

group A and provide a detailed analysis of group B, using (a) the joint regression model, (b) our computational algorithm and (c) the global search strategy, as described above.
First we investigated whether or not a saturated mean was required. Like Pourahmadi (1999) we used two cubic polynomials for modelling the autoregressive and innovation parameters. Figure 1 shows that, when p is varied, BIC$(p, 3, 3)$ takes its minimum at $p = 8$ and not at the saturated mean $p = 10$. The BIC values between $p = 5$ and $p = 10$ are very similar, suggesting that a saturated mean is un-necessary. Secondly, we studied the effect on the other parameters of varying p in the $(p, 3, 3)$ model. Figure 2 illustrates the effects on the innovation variances. One can see that a sub-optimal choice of p influences the estimation of the other parameters - contrary to much current statistical thinking. Thirdly we identified the optimal model as BIC$(8, 4, 3) = 71.73$ and not BIC$(10, 3, 3) = 71.89$, as claimed by Pourahmadi. The difference in BIC values is small, in this case, but the structures implied by the two models are rather different - regressogram-based inference (with a saturated mean model) having failed to correctly identify q, the degree of the polynomial for the autoregressive coefficients. In further work, we investigated model mis-specification by systematically over-fitting the optimal model ie, by saturating one, two and three dimensions in turn. These results show that mis-specification of the innovation variances is most serious.
Analysis of Group B, in which simple cubic trends are absent from the regressograms, showed that the regressogram approach was more suboptimal than in Group A, that our computational algorithm always converged, and that our search strategy identified the global optimum. Details will appear in the main paper.

5 Discussion

We undertook this work to improve our understanding of joint mean-covariance modelling in the analysis of longitudinal studies. We have demonstrated that the use of a saturated mean model in the data analyzed by Pourahmadi is not optimal and that the use of the regressograms can lead to mis-specified models. Moreover, it is clear that in some circumstances none of the components of the optimal triple may be identified correctly by regressogram inference. The problems encountered are not confined to the model class investigated, but, in principle, are likely to arise, whenever the observed information matrix is not block diagonal. Accordingly, we cannot endorse their routine use at this time, but advocate instead the BIC-based profile search method proposed above.

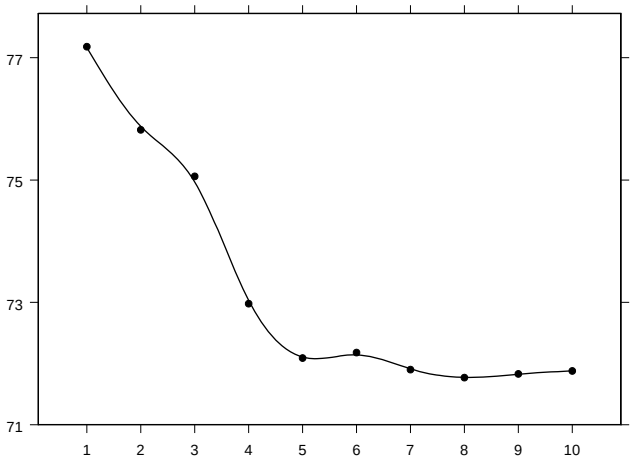


FIGURE 1. *BIC*$(p, 3, 3)$ *versus* p

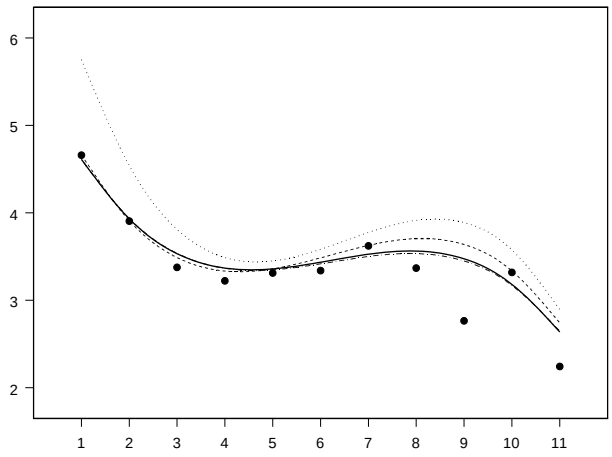


FIGURE 2. *Innovation variances versus time*

References

Barndorff-Nielson, O. (1991). Likelihood theory. In *Statistical Theory and Modelling*. Ed. D.V. Hinkley, N. Reid, and J. Snell, London: Chapman & Hall, 232-264.

Diggle, P.J., Liang, K.Y., and Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

Goldstein H., Healy M.J.R., and Rasbash J. (1994) . Multilevel time series models with applications to repeated measures data. *Statistics in Medicine*, **13**, 1643–55.

Kenward, M.G. (1987). A method for comparing profiles of repeated measurements. *Applied Statistics*, **36**, 296–308.

Pan J.X. and MacKenzie G. (2003). On modelling mean-covariance structures in longitudinal studies *Biometrika*. (in press).

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, **86**, 677–90.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, **87**, 425–35.

Rao, C.R. (1987). Prediction of future observations in growth curve models (with Discussion). *Statistical Science*, **2**, 434–471.

Hierarchical Modeling of Health Services Outcome and Resource Use: Issues in Hospital Performance Comparison Studies

Ying MacNab[1], Zhenguo Qiu[1], Paul Gustafson[2], Charmaine Dean[3], Shoo Lee[2], and Arne Ohlsson[4]

[1] Centre for Healthcare Innovation and Improvement British Columbia Institute for Children's and Women's Health University of British Columbia 4480 Oak Street, Rm E-414, Vancouver, B.C., Canada V6H 3V4
[2] Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada, V6T 1Z2
[3] Department of Statistics and Actuarial Science, Simon Fraser University, Burnday, BC, Canada V5A 1S6
[4] Department of Pediatrics, University of Toronto, Toronto, Ontario, Canada M5G 1X5

Abstract: This presentation begins with a brief introduction of the Neonatal Health Services in Canada Project. The study, currently funded by Canadian Institute for Health Research, aims to examine the impact of geography, local health access and health care systems on variations in outcomes and resource use in neonatal intensive care units (NICUs) across Canada. The assessment of NICU outcomes and resource utilization is discussed from the viewpoint of statistical modeling. As an illustration, we present an in-depth analysis of neonatal mortality variation among 17 Canadian NICUs, with covariates available at both the patient and NICU levels. We describe the use of hierarchical Bayesian models for systematically exploring outcome heterogeneity between NICUs and discuss statistical issues relating to multilevel modeling, Bayesian computation via MCMC, analysis involving outlying observations, and estimation of multi-level effects.

Keywords: Hierarchical logistic regression model; Markov Chain Monte Carlo; Random effects; Mortality in neonatal intensive care unit; Institutional comparison of outcomes.

Overview

Recent literature illustrates the potential of hierarchical Bayesian methodology as a general framework of substantial flexibility for multilevel analysis of outcome variations within the context of institutional comparison and provider (i.e. physician, hospital, teacher, school) profiling. Studies to date have largely focused on institutional comparison or provider profiling through the development of risk-adjusted performance indicators, for

example, risk-adjusted mortality rates or performance score, among institutions or providers. A common goal of these studies was the ranking of adjusted performance indicator or the identification of 'under-performed' provider(s).

In this presentation, we discuss more general issues surrounding quantitative comparisons of institutional outcomes and development of analytic strategies for comprehensive analysis of multilevel health services data in order to provide detailed information about important sources of outcome variation. We demonstrate the extended potentials of the Bayesian hierarchical modeling as a general strategy for systematically evaluating response-covariate associations at each level of the hierarchy, examining cross-level interaction, quantifying residual variance or attribution of unexplained variability, and deriving various types of risk-adjusted and risk-specific inter-provider comparisons. We present an analysis of neonatal mortality variation among 17 neonatal intensive care units (NICU) across Canada and engage an extensive discussion on relevant issues that are important for institutional comparison, provider profiling, and quality improvement efforts.

In the past decade, the development of Markov chain Monte Carlo (MCMC) methods has made it possible to implement full Bayesian inference in multilevel modeling. The availability of various MCMC methods greatly extends the potential of Bayesian hierarchical models as a general framework for comprehensive analysis of systematic variation arising from various sources and for adequate assessment of model uncertainty and estimation precision. In this study, we illustrate Bayesian analysis of multilevel data and the implementation of MCMC computation. We emphasize particular contributions of Bayesian hierarchical modelling of population heterogeneity in institutional comparison studies, and discuss statistical issues relating to Bayesian computation and inference, estimation of multi-level effects, and analysis involving high-leverage observations.

Smooth Regression Coefficient Surfaces

Brian D. Marx[1] and Paul H.C. Eilers[2]

[1] Department of Experimental Statistics, Louisiana State University, Baton Rouge, LA 70803 USA (bmarx@lsu.edu)
[2] Paul H. C. Eilers, Department of Medical Statistics, Leiden University Medical Center, 2300 RA, Leiden, The Netherlands, (p.eilers@lumc.nl)

Abstract: We propose a general approach to regression on "images" that can pose severe challenges to standard statistical methods. The main contribution of this work is to build a two-dimensional coefficient surface that allows for interactive features across the indexing plane of the regressor array. We aim to use the estimated coefficient surface for reliable (scalar) prediction. We assume that the coefficients are smooth along both indices. We present a rather straight-forward and rich extension of penalized signal regression (Marx and Eilers, 1999) using penalized B-spline tensor products, where appropriate difference penalties are placed on the rows and columns of the tensor product coefficients. Our methods are are grounded in standard penalized regression, thus cross-validation, effective dimension and other diagnostics are accessible. Further the model is easily transplanted into the generalized linear model framework.

Keywords: Multivariate calibration; P-splines; Signal regression; Tensor product.

1 Introduction

Consider fluorescence spectroscopy experiments: for each response, there are thousands of regressors arranged in a two-dimensional array (along emission and excitation axes). The problem is inherently ill-posed, as often the number of samples in the training data is far less than the number of array elements. Bro (1998) presented sugar process data that consisted of several scalar quality measurement responses (e.g. ash content and color) and regressor information that consisted of emission spectra (at 571 wavelengths) across seven excitation wavelengths. This yields an array of 3997 regressors, but there are only $m = 265$ training samples. Such data structure is not specific to chemometric applications: one can imagine medical images for several (hundred) patients. The image can be viewed as a regressor surface (e.g. 64×64 grey-scale pixels). The response may be a binary indicator of presence or absence of some tumor feature, and modelling such data may further require structure of the generalized linear model. Eilers and Marx (2003) had success on a similar, simpler, problem: one using (functional) spectra regressors, which could be viewed as a "very narrow

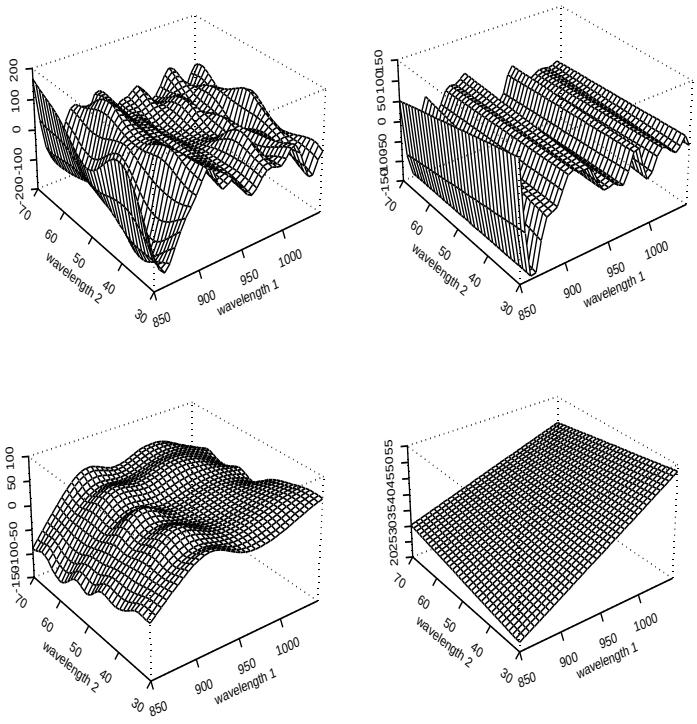


FIGURE 1. *Examples of surfaces that can be generated from tensor products when constraining roughness in two dimensions. Effective dimension (upper, left:74); (upper, right: 34); (lower, left: 23); (lower, right: 5).*

image". To provide an idea of how a surface may look Figure 1 provides examples of coefficient surfaces using tensor products. The upper, left panel displays a surface constructed from essentially unpenalized tensor products, whereas the lower, right surface displays the limiting plane resulting from large second order penalties on every row and column of tensor products. The other two figures have a mixture of a low penalty on one axis and a high penalty on the other.

2 Tensor Product *B*-splines in a Nutshell

Eilers and Marx (2003) provided an overview of tensor products. Figure 2 displays a portion of a full tensor product basis. As seen, tensor product B-splines exist in, say, the $v \times t$ plane. There are knots selected on an equally-spaced grid, carving out the plane into subrectangles. A tensor

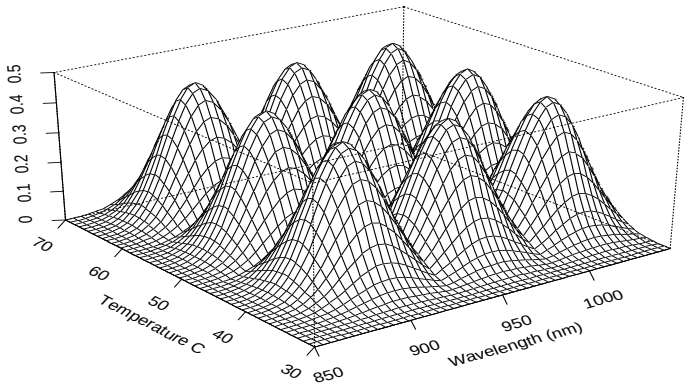


FIGURE 2. *A portion of a full tensor product basis.*

product $B_r(v)\breve{B}_s(t)$ is positive in the rectangular region defined by the knots $R = [v_r,\ v_{r+q_v+2}] \times [t_s,\ t_{s+q_t+2}]$ or on a support of spanned by $(q_v + 2) \times (q_t + 2)$ knots, where q is the degree of the B-spline. Specifically, each tensor product can be indexed by one of $(n_v \times n_t)$ knot pairs ($r = 1,\ \ldots,\ n_v$ and $s = 1,\ \ldots\ , n_t$) and

$$B_r(v)\breve{B}_s(t) \geq 0 \text{ for all } v,\ t \in R; \quad \text{else zero} \tag{1}$$

Some technical details follow. We choose to divide the domain v (wavelength 1): $v_{\min}$ to $v_{\max}$ into n' equal intervals, using $n' + 1$ interior knots. Taking each boundary into consideration, a complete basis needs $n' + 2q + 1$ total knots. Denote the knots as: $v_1,\ \ldots,\ v_{n'+2q+1}$. The total number of B-splines on the axis is $n = n' + q$. For indexing purposes it is convenient to associate each B-spline, $B_r(v)$ with exactly one of the (first) $k = 1,\ \ldots,\ n$ knots. A similar division of the t axis (wavelength 2) is made for $\breve{B}_r(t)$, also using equally-spaced knots, but possibly using a different q, n' or p. Denote $\Gamma_{n_v \times n_t} = [\gamma_{rs}]$ as the matrix of unknown tensor product B-spline coefficients. For given knot grid, a very flexible surface can be approximated at each of the digitized spectra surface coordinates (v_j, t_k)

$(j = 1, \cdots,\ p_1;\ k = 1, \cdots,\ p_2)$ by

$$\alpha(v_j, t_k) = \sum_{r=1}^{n_v} \sum_{s=1}^{n_t} B_r(v_j) \breve{B}_s(t_k) \gamma_{rs}, \tag{2}$$

where $r = 1,\ \ldots,\ n_v$ and $s = 1,\ \ldots,\ n_t$. The system of equations is of order $n_v n_t$. The $(p_1 p_2) \times 2$ matrix of regressor locations is $\jmath = (v \otimes \mathbf{1}_{p_2},\ \mathbf{1}_{p_1} \otimes t)$. The matrix $\mathbf{B}_v$ and $\breve{\mathbf{B}}_t$ are evaluated at the first and second column of $\jmath$, respectively. It is computationally efficient (by avoiding looping) to reexpress the surface in matrix notation as $\alpha(v, t) = \mathbf{B}\gamma$, where $\gamma = \text{vec}(\Gamma)$ (of length $n_v n_t$), and

$$\mathbf{B} = \mathbf{B}_v \square \breve{\mathbf{B}}_t = (\mathbf{B}_v \otimes \mathbf{1}'_{n_t}) \odot (\mathbf{1}'_{n_v} \otimes \breve{\mathbf{B}}_t). \tag{3}$$

The symbols $\otimes$ and $\odot$ denote Kronecker product and elementwise multiplication of matrices, respectively. The matrix $\mathbf{B}$ is of dimension $(p_1 p_2) \times (n_v n_t)$, i.e. the $p_1 p_2$ (digitized) surface is (initially) projected onto a $n_v n_t$ smooth dimensional surface. Penalized estimation of γ and its use with regressor surfaces are next discussed.

3 Penalized Two-Dimensional Coefficient Surfaces

Given the ith regressor matrix $X_i = [x_{ijk}]$ of dimension $p_1 \times p_2$ $(i = 1, \ldots,\ m;\ j = 1,\ \ldots,\ p_1;\ k = 1, \ldots,\ p_2)$ and coefficient surface $\alpha(v, t)$, express the mean

$$\mu_i = \alpha_0 + \sum_{j=1}^{p_1} \sum_{k=1}^{p_2} x_{ijk} \alpha(v_j, t_k). \tag{4}$$

Using tensor product B-splines, (2) can be substituted into (4) yielding

$$\mu_i - \alpha_0 = \sum_{j=1}^{p_1} \sum_{k=1}^{p_2} x_{ijk} \sum_{r=1}^{n_v} \sum_{s=1}^{n_t} B_r(v_j) \breve{B}_s(t_k) \gamma_{rs} = \mathbf{x}'_i \mathbf{B} \gamma, \tag{5}$$

where $\mathbf{x}'_i = \text{vec}(X_i)$. We aim to find a practical solution to minimize $Q(\alpha_0, \gamma) = |y - \alpha_0 - \mathbf{M}\gamma|^2$, where $\mathbf{X}$ is of dimension $m \times (p_1 p_2)$ and $\mathbf{M} = \mathbf{XB}$. The use of tensor product B-splines does reduce the dimension of estimation, but there are still $n_v n_t + 1$ unknown parameters. For even moderately complex surfaces, ill-posed estimation problems can arise, as it may be necessary to increase the number on knots on the grid to allow enough flexibility.

In the spirit of P-splines (Eilers and Marx, 1996), appropriate penalties can be put on γ and thus regularize estimation. A separate difference penalty

is attached to each of the rows and each of the columns of Γ. The objective function is now modified to minimize

$$Q^\star(\alpha_0, \gamma) = |y - \alpha_0 - \mathbf{M}\gamma|^2 + \lambda_v |P_v \gamma|^2 + \lambda_t |P_t \gamma|^2 + \lambda_0 |\gamma|^2. \qquad (6)$$

The last term in (6) is an overall ridge penalty. Indexing can quickly get out of hand and the penalties are most compactly represented in matrix notation as: $P_v = (D_d' D_d) \otimes I_{n_t}$ and $P_t = I_{n_v} \otimes (D_d' D_d)$, where I denotes the identity matrix. Although it is not reflected in the notation, the order of the row penalty (d_v) can be different from that of the column penalty (d_t). Much like the PSR approach, the difference penalties ensure that adjacent coefficients with in the same row (or a column) do not differ too much from each other. The penalties can continuously regulate roughness through the nonnegative λ_v, λ_t and λ_0.
The explicit P-spline solution for (6) is

$$\hat{\gamma} = (\mathbf{M}^{\star'}\mathbf{M}^\star + \lambda_v P_v^\star + \lambda_t P_t^\star + \lambda_0 I^\star)^{-1} \mathbf{M}^{\star'} y,$$

with $\mathbf{M}^\star = (1_m | \mathbf{M})$, $P^\star = (0|P)$, and $I^\star = \mathrm{diag}(0, I_{n_v n_t})$. The predicted values are $\hat{y} = \mathbf{M}^\star(\hat{\alpha}_0, \hat{\gamma}')'$. The "hat" matrix is $H = [h_{ii'}] = \mathbf{M}^\star(\mathbf{M}^{\star'}\mathbf{M}^\star + \lambda_v P_v^\star + \lambda_t P_t^\star + \lambda_0 I^\star)^{-1}\mathbf{M}^{\star'}$. Given $\hat{y}$ and the diagonal of H, leave-one-out cross-validation standard error of prediction can be quickly calculated:

$$\mathrm{CV}(\lambda_v, \lambda_t, \lambda_0) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\frac{y_i - \hat{y}_i}{1 - h_{ii}}\right)^2}.$$

The optimal $(\lambda_v, \lambda_t, \lambda_0)$ can be found, e.g. using a search to minimize CV. The dimension of the estimated coefficient surface can be approximated by $\mathrm{trace}(H)$, and the error variance component estimated by

$$\hat{\sigma}^2 = \frac{|y - \hat{y}|^2}{m - \mathrm{trace}(H)}.$$

For computational purposes, it is worth mentioning that the explicit solution for (6) can be found efficiently through data augmentation tricks, i.e. $(\hat{\alpha}_0, \hat{\gamma}')' = (\mathbf{M}_+^{\star'}\mathbf{M}_+^\star)^{-1}\mathbf{M}_+^{\star'} y_+$, where

$$\mathbf{M}_+^\star = \begin{bmatrix} 1 & \mathbf{M} \\ 0 & \sqrt{\lambda_v}(D_d \otimes I_{n_t}) \\ 0 & \sqrt{\lambda_t}(I_{n_v} \otimes D_d) \\ 0 & \sqrt{\lambda_0} I_{n_v n_t} \end{bmatrix} \quad \text{and} \quad y_+ = \begin{bmatrix} y \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Two-dimensional spectroscopic data often have a parallelogram shape leaving unsupported regions in $v \times t$: however the penalty automatically remedies this problem. Since the coefficient surface model is grounded in (penalized) least squares regression techniques, the methodology can be easily

extended into the generalized linear model, e.g. using the penalized scoring algorithm with binomial or Poisson responses. Optimization of the tuning parameters can then be identified by minimizing and information criteria, a simple function of deviance and effective dimension. Lastly, simpler models can be investigated that use varying coefficient structure.

References

Bro, R. (1998). Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis. *Chemometrics and Intelligent Laboratory Systems*, **46**, 133–147.

Eilers, P.H.C. and Marx, B.D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems*. (in press).

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

Marx, B.D. and Eilers, P.H.C. (1999). Generalized linear regression on sampled signals and curves: a P-spline approach. *Technometrics*, **41**, 1–13.

Modelling some Experiments Carried out in Incomplete Split-Block-Plot Designs

Iwona Mejza[1] and Katarzyna Ambroży[1]

[1] Department of Mathematical and Statistical Methods, Agricultural University, Wojska Polskiego 28,60-637 Poznań, Poland

Abstract: Model building of data obtained from three-or-more-factor experiments carried out in incomplete split-block-plot design is presented. In the modelling the structure of an experimental material and a four-step randomization scheme are taken into account. In the construction method some efficiency-balanced designs are considered. With respect to the analysis of the obtained randomization model with six strata the approach typical to the multistratum experiments with orthogonal block structure is adopted.

Keywords: General balance; Mixed design; Split-block design; Split-plot design.

1 Introduction

Some experimental designs used in agricultural research for three-or-more factor experiments are extensions of either a split-plot or a split-block design (cf. Gomez and Gomez, 1984). The considered here design is the extension of the split-block design in which the intersection plot is divided into subplots to accommodate a third factor. Another term of the design is the strip-split-plot design (cf. Gomez and Gomez, 1984). We can note that it is a mixed design of the split-block design for two first factors and the split-plot design with treatment combinations of the first two factors and the third factor.,
In the paper we consider a situation when the split-block-plot (SBP) design is incomplete with respect to (w.r.t.) one or more factors. In the construction method some efficiency balanced (EB) designs, in particular balanced incomplete block (BIB) designs are taken into account (cf. Cochran and Cox, 1957, Caliński and Kageyama, 2000).

2 Assumptions and Notation

Let us consider a three-factor experiment of a SBP 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.
We assume that experimental material is divided into b blocks. Every block forms a row−column design with k_1 rows and k_2 columns. Then each intersection plot (called also whole plot) is divided into k_3 subplots. So, the number of observations is equal to n $(= bk_1k_2k_3)$. Here the rows correspond to the levels of the factor A, termed also as row treatments, the columns correspond to the levels of the factor B, called also column treatments, and the subplots are to accommodate the levels of the factor C termed as subplot treatments.
It can be noted there are four plot sizes (the row, the column, the whole plot and the subplot), so there are four levels of precision with which the effects of the various factors are estimated. The highest level corresponds to the subplot factor and its interactions with other factors. The precision is strictly connected with efficiency of the estimation of the contrasts (comparisons) of the treatment combinations. It is well known that the efficiency is the highest (full efficiency) in the complete (in particular orthogonal, if it exists) design.
We consider a situation that SBP design with $k_1 \leq s$, $k_2 \leq t$, $k_3 \leq w$ is incomplete w.r.t. one factor (A or B or C) only, two factors only or is incomplete w.r.t. all the factors. It means those first factors, A and B, can be arranged as in an incomplete split-block design (cf. Hering and Mejza S., 1997, Mejza I., 1998) and the third factor, C, can be arranged as in an incomplete split-plot design (cf. Mejza and Mejza, 1984).

3 Linear Model

Since all units have to be randomized before they enter the experiment, we perform the four-step randomization. Let us note that first three steps of the randomization connected with the blocks, the rows and the columns are strictly the same as in the split-block design whereas the fourth step connected with the subplots takes place as in the split-plot design. So, this mixed process of randomization leads to a randomization model with five main strata (without zero stratum connected with mean of an experiment only). This model is of the form:
where $\mathbf{\Delta}'$ is a known design matrix for v treatment combinations, and τ $(v \times 1)$ is the vector of fixed treatment combination effects. According to the orthogonal block structure of the SBP designs, the dispersion matrix $\mathbf{V}(\gamma)$ can be expressed by $\{\mathbf{V}(\gamma) = \sum_{f=0}^{5} \gamma_f \mathbf{P}_f)\}$ where $\gamma_f \geq 0$ and $\{\mathbf{P}_f\}$ are a family of known pairwise orthogonal projectors adding up to the identity matrix (cf. Houtman and Speed, 1983). The range space $\Re\{\mathbf{P}_f\}$ of $\mathbf{P}_f$, $f = 0, 1, \ldots, 5$, is termed the f-th stratum of the model and γ_f are unknown strata variances. This model will be analyzed using the methods developed for multistratum experiments. So, we have zero stratum (0) generated by

the vector of ones, inter-block stratum (1), inter-row (within the block) stratum (2), inter-column (within the block) stratum (3), inter-whole plot (within the block) stratum (4) and inter-subplot (within the whole plot) stratum (5).
Let us assume that the treatment combinations are ordered lexicographically. It is well known that statistical properties of the design are strictly connected with the algebraic properties of the stratum information matrices for the treatment combinations $\mathbf{A}_f$, $f = 1, ..., 5$. It is useful to express them by the known from the theory of block designs matrices $\mathbf{C}_f$, $f = 0, 1, 2, 3, 4$. They have the forms:

$$\mathbf{C}_0 = \mathbf{r}^\delta - n^{-1}\mathbf{r}\mathbf{r}',$$

$$\mathbf{C}_1 = \mathbf{r}^\delta - (k_1k_2k_3)^{-1}\mathbf{N}_1\mathbf{N}_1', \qquad \mathbf{C}_2 = \mathbf{r}^\delta - (k_2k_3)^{-1}\mathbf{N}_2\mathbf{N}_2',$$

$$\mathbf{C}_3 = \mathbf{r}^\delta - (k_1k_3)^{-1}\mathbf{N}_3\mathbf{N}_3', \qquad \mathbf{C}_4 = \mathbf{r}^\delta - k_3^{-1}\mathbf{N}_4\mathbf{N}_4',$$

where $\mathbf{N}_1$, $\mathbf{N}_2$, $\mathbf{N}_3$, $\mathbf{N}_4$ are treatments vs. blocks -, treatments vs. rows -, treatments vs. columns - and treatments vs. whole plots incidence matrices, respectively, $\mathbf{r} = \mathbf{N}_1\mathbf{1}_b = \mathbf{N}_2\mathbf{1}_{bk_1} = \mathbf{N}_3\mathbf{1}_{bk_2} = \mathbf{N}_4\mathbf{1}_{bk_1k_2}$ is the vector of replications of the treatment combinations, $\mathbf{r}^\delta = diag(r_1, r_2, ..., r_v)$ and $\mathbf{1}_x$ is the x-dimensional vector of ones.
These matrices are $\mathbf{C}$-matrices of the associated orthogonal design, the block design, the row design with rows as blocks, the column design with columns as blocks and the whole plot design with whole plots as blocks, respectively. Then the information matrices $\mathbf{A}_f$ can be expressed as follows:

$$\mathbf{A}_1 = \mathbf{C}_0 - \mathbf{C}_1, \quad \mathbf{A}_2 = \mathbf{C}_1 - \mathbf{C}_2, \quad \mathbf{A}_3 = \mathbf{C}_1 - \mathbf{C}_3$$

$$\mathbf{A}_4 = \mathbf{C}_2 + \mathbf{C}_3 - \mathbf{C}_1 - \mathbf{C}_4, \quad \mathbf{A}_5 = \mathbf{C}_4.$$

Let ϵ_{fh} denote an eigenvalue of the matrix $\mathbf{A}_f$ w.r.t. $\mathbf{r}^\delta$, corresponding also to an eigenvector $\mathbf{s}_h$, $f = 1, ..., 5, h = 1, 2, ..., v$. Since $\mathbf{A}_f\mathbf{1}_v = \mathbf{0}$, the last eigenvector $\mathbf{s}_h$ may be chosen as $n^{\frac{1}{2}}\mathbf{1}_v$.
We can note that $\mathbf{p}_h = \mathbf{r}^\delta\mathbf{s}_h$, defines a (basic) contrast $\mathbf{p}_h'\tau$, $h = 1, 2, ..., v - 1$. These contrasts are strictly connected with the comparisons among the main effects of the considered factors and interaction effects between them. Stratum efficiency factors of the considered SBP designs w.r.t. these contrasts are expressed by the eigenvalues ϵ_{fh}, $f = 1, ..., 4, h = 1, 2, ..., v - 1$.

4 Some Case of the Split-Block-Plot Design

It is convenient to 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 efficiency α in the f - th stratum. In other words, we say that the design is $M_f\{q,\alpha\}$ - balanced. Particularly, for $\alpha = 1$ the design is $M_f\{q,1\}$ - orthogonal.
Let us consider the SBP design incomplete w.r.t. the row and column treatments and complete w.r.t. the subplot treatments.
Let $\mathbf{N}_A(s \times b)$ and $\mathbf{N}_B(t \times b)$ be incidence matrices for the row and the column treatments w.r.t. the blocks, respectively. Then we have: $\mathbf{N}_1 = \mathbf{N}_A \otimes \mathbf{N}_B \otimes \mathbf{1}_w$ and $\mathbf{r} = \mathbf{r}_A \otimes \mathbf{r}_B \otimes \mathbf{1}_w$, where $\mathbf{r}_A = \mathbf{N}_A\mathbf{1}_A$, $\mathbf{r}_B = \mathbf{N}_B\mathbf{1}_B$,

$$\begin{aligned}
\mathbf{C}_0 &= \mathbf{r}_A^\delta \otimes \mathbf{r}_B^\delta \otimes \mathbf{I}_w - n^{-1}\mathbf{r}_A\mathbf{r}'_A \otimes \mathbf{r}_B\mathbf{r}'_B \otimes \mathbf{J}_w,\\
\mathbf{C}_1 &= \mathbf{r}_A^\delta \otimes \mathbf{r}_B^\delta \otimes \mathbf{I}_w - (k_1k_2w)^{-1}(\mathbf{N}_A\mathbf{N}'_A \otimes \mathbf{N}_B\mathbf{N}'_B \otimes \mathbf{J}_w),\\
\mathbf{C}_2 &= \mathbf{r}_A^\delta \otimes \mathbf{r}_B^\delta \otimes \mathbf{I}_w - (k_2w)^{-1}(\mathbf{r}_A^\delta \otimes \mathbf{N}_B\mathbf{N}'_B \otimes \mathbf{J}_w),\\
\mathbf{C}_3 &= \mathbf{r}_A^\delta \otimes \mathbf{r}_B^\delta \otimes \mathbf{I}_w - (k_1w)^{-1}(\mathbf{N}_A\mathbf{N}'_A \otimes \mathbf{r}_B^\delta \otimes \mathbf{J}_w),\\
\mathbf{C}_4 &= \mathbf{r}_A^\delta \otimes \mathbf{r}_B^\delta \otimes \mathbf{I}_w - w^{-1}(\mathbf{r}_A^\delta \otimes \mathbf{r}_B^\delta \otimes \mathbf{J}_w),
\end{aligned}$$

where $\mathbf{r}_A^\delta = diag(r_1^A, r_2^A, ..., r_s^A)'$ and $\mathbf{r}_B^\delta = diag(r_1^B, r_2^B, ..., r_t^B)'$.
Let $\mathbf{K}^{(A)} = \{h : h = 1, 2, ..., s-1\}$ and $\mathbf{K}^{(B)} = \{m : m = 1, 2, ..., t-1\}$ and let

$$\mathbf{C}_A = \mathbf{r}_A^\delta - k_1^{-1}\mathbf{N}_A\mathbf{N}'_A \text{ with eigenvalues } \mu_1, \mu_2, ..., \mu_s \text{ w.r.t. } \mathbf{r}_A^\delta,$$

$$\mathbf{C}_B = \mathbf{r}_B^\delta - k_2^{-1}\mathbf{N}_B\mathbf{N}'_B \text{ with eigenvalues } \xi_1, \xi_2, ..., \xi_t \text{ w.r.t. } \mathbf{r}_B^\delta.$$

Following algebraic properties of the $\mathbf{C}_f$, $f = 0, 1, ..., 4$ given above and the structures of the matrices $\mathbf{C}_A$ and $\mathbf{C}_B$ we have:

Corollary 1.
The considered incomplete SBP design is:
$A_1\{1, 1-\mu_h\}$ - balanced, $h \in \mathbf{K}^{(A)}$, $B_1\{1, 1-\xi_m\}$ - balanced, $m \in \mathbf{K}^{(B)}$,
$(A \times B)_1\{1, (1-\mu_h)(1-\xi_m)\}$ - balanced, $h \in \mathbf{K}^{(A)}$, $m \in \mathbf{K}^{(B)}$,
$A_2\{1, \mu_h\}$ - balanced, $h \in \mathbf{K}^{(A)}$,
$(A \times B)_2\{1, \mu_h(1-\xi_m)\}$ - balanced, $h \in \mathbf{K}^{(A)}$, $m \in \mathbf{K}^{(B)}$,
$B_3\{1, \xi_m\}$ - balanced, $m \in \mathbf{K}^{(B)}$,
$(A \times B)_3\{1, (1-\mu_h)\xi_m\}$ - balanced, $h \in \mathbf{K}^{(A)}$, $m \in \mathbf{K}^{(B)}$,
$(A \times B)_4\{1, \mu_h\xi_m\}$ - balanced, $h \in \mathbf{K}^{(A)}$, $m \in \mathbf{K}^{(B)}$,
$C_5\{w-1, 1\}$ - orthogonal,
$(A \times C)_5\{(s-1)(w-1), 1\}$ - orthogonal,
$(B \times C)_5\{(t-1)(w-1), 1\}$ - orthogonal
and $(A \times B \times C)_5\{(s-1)(t-1)(w-1), 1\}$ - orthogonal.
We can notice that all contrasts connected with main effects of the factor C and with its interaction effects with other factors are estimated with full efficiency in the inter-subplot stratum. Other contrasts are estimable in two strata (between main effects of A and B) and four strata (between interaction effects $A \times B$).
If EB (in particular BIB) designs are considered as generating designs for the row and column treatments the number of efficiency classes reduces.

Let $\mathbf{C}_A = \mu(\mathbf{r}_A^{\delta} - (bk_1)^{-1}\mathbf{r}_A\mathbf{r}'_A)$ and $\mathbf{C}_B = \xi(\mathbf{r}_B^{\delta} - (bk_2)^{-1}\mathbf{r}_B\mathbf{r}'_B)$ be $\mathbf{C}$-matrices of the EB (in particular BIB) designs and let
$\mu = [\ bk_1 - k_1^{-1}tr(\mathbf{N}_A\mathbf{N}'_A)]\ /[\ bk_1 - (bk_1)^{-1}\mathbf{r}'_A\mathbf{r}_A)]$ and
$\xi = [\ bk_2 - k_2^{-1}tr(\mathbf{N}_B\mathbf{N}'_B)]\ /[\ bk_2 - (bk_2)^{-1}\mathbf{r}'_B\mathbf{r}_B]$ be their eigenvalues, respectively (e.g. Caliński and Kageyama, 2000).
Then we can express:

Corollary 2.
The considered SBP design with EB (in particular BIB) design for the row and column treatments is:
$A_1\{s-1, 1-\mu\}$ - balanced, $B_1\{t-1, 1-\xi\}$ - balanced,
$(A\times B)_1\{(s-1)(t-1), (1-\mu)(1-\xi)\}$ - balanced,
$A_2\{s-1, \mu\}$ - balanced, $(A\times B)_2\{(s-1)(t-1), \mu(1-\xi)\}$ - balanced,
$B_3\{t-1, \xi\}$ - balanced, $(A\times B)_3\{(s-1)(t-1), (1-\mu)\xi\}$ - balanced,
$(A\times B)_4\{(s-1)(t-1), \mu\xi\}$ - balanced, $C_5\{w-1, 1\}$ - orthogonal,
$(A\times C)_5\{(s-1)(w-1), 1\}$ - orthogonal, $(B\times C)_5\{(t-1)(w-1), 1\}$ - orthogonal and $(A\times B\times C)_5\{(s-1)(t-1)(w-1), 1\}$ - orthogonal.

Acknowledgments: The work was partially supported by KBN grant no 3 P06A 017 22.

References

Caliński, T. and Kageyama, S. (2000). *Block Designs: A Randomization Approach. Vol. 1: Analysis*. Lecture Notes in Statistics, **150**. New York: Springer-Verlag.

Cochran, W.G. and Cox, G.M. (1957). *Experimental designs*. New York: Wiley.

Gomez, K.A. and Gomez, A.A. (1984). *Statistical Procedures for Agricultural Research*. New York: Wiley.

Hering, F. and Mejza, S. (1997). Incomplete split-block designs. *Biometrical Journal*, **39**, 227–238.

Houtman, A.M. and Speed, T.P. (1983). Balance in designed experiments with orthogonal block structure. *Annals in Statistics*, **11**, 1069–1085.

Mejza, I. (1998). Characterisation of certain split-block designs with a control. *Biometrical Journal*, **40**, 627–639.

Mejza, I. and Mejza, S. (1984). Incomplete split-plot designs. *Statistics and Probability Letters*, **2**, 327–332.

Effect of the Use of Credit Cards on Italian Families' Liquidity: an Empirical Evaluation

Andrea Mercatanti[1]

[1] Department of Statistics and Applied Mathematics, University of Pisa,Via C.Ridolfi 10, 56124, Pisa, Italy. mercatan@ec.unipi.it

Abstract: The paper constitutes an application of causal inference methods to a microeconomic dataset. The economic question is to evaluate the effect of the use of credit cards on the Italian families' liquidity. In order to take into account the self-selection of the units to the treatment, the Instrumental Variables method (Imbens and Angrist, 1994; Angrist et al, 1996) and the Two Stage Model (Heckman, 1978; 1979) are applied. The final result is a negative and significant causal effect of credit cards on the minimal amount of cash held by Italian families.

Keywords: Causal inference; Instrumental variables; Two-stage model.

1 Introduction

The aim of the paper is to quantify the effect of the use of credit cards on the minimal amount of cash that Italian families held at home and under which a withdrawal becomes necessary. The main justification is in the fact that credit cards can be considered as close substitutes for cash in small amount payments. Consequently a significant effect of credit cards on the amount of cash families need for the everyday life would contribute in explaining the different liquidity choices induce by the use of alternative payment instruments and eventually in suggesting further directions of research. But the analysis is complicated by the fact that credit cards can act, other than as substitute of money, as a mean of encouragement to buy and of consumer credit; and each of these potential causes can have effects of different directions on liquidity. Moreover, a first descriptive analysis shows a self-selection of the units to the treatment. Then an evaluation of the overall effect required to take in to account all the potential causes and this is possible only by an appropriate use of causal inference methods.
The data used in this application are from the sample survey "The Italian families conditions in 1995 *(I bilanci delle famiglie italiane nell'anno 1995)*" run by the Italian Central Bank *(Banca d'Italia)*. The statistical unit is the Italian family, and the size of the sample is 6586 units.

2 Methodology

The analysis is performed by using statistical methods for causal inference and it is based on the concept of potential outcomes. Following this ap-

TABLE 1. *Conditional relative frequencies of the pre-treatment variable "Householder position".*

Householder position:	$D_i = 0$	$D_i = 1$
blue-collar	0.190	0.083
white-collar	0.174	0.401
manager, high officials	0.009	0.066
professional man, entrepreneur	0.140	0.246
unemployed	0.026	0.011
housewife, retired	0.461	0.193

proach, causal effects are defined by comparing average potential outcomes that would have been observed under different treatments: *Average Treatment Effect,* A.T.E. (Holland, 1986). In this application the outcome is the minimal amount of cash held by a family (Y_i), and the treatment is a binary indicator of the presence of at least a credit card holder in the family (D_i). Under the assumption of random assignment to treatment, A.T.E. can be estimated simply by comparing the average outcomes in the group of treated and non-treated. But a first descriptive analysis shows that the pre-treatment variables are unbalanced in the subsamples defined by the treatment (Table 1 shows the results for the pre-treatment variable "Householder position"). This is in contrast to the assumption of randomization and justifies the use of appropriate causal inference methods for taking into account the self-selection of the units to the treatment.
To this purpose, two different statistical methods will be used: the method based on Instrumental Variables, I.V. (Imbens and Angrist, 1994; Angrist et al., 1996), and the *"Two Stages Model"* (Heckman, 1978, 1979). In first analysis the non-parametric and less restrictive I.V. method is applied. Angrist et al (1996) showed that, by introducing an instrumental variable having the role of "random assignment to treatment", Z_i, and under weaker albeit crucial assumptions, it is possible to identify and estimate causal effects without relying on distributional assumptions. More precisely the I.V. estimate, $\hat{\beta}_{IV}$, in the regression

$$Y_i = \alpha_{IV} + \beta_{IV} D_i + \varepsilon_i$$

identifies the treatment effect, for the group of people complying with the assignment to treatment (Local Average Treatment Effect: L.A.T.E.). This result can be extended to the whole population under the further assumption that the treatment effect for non-compliers is equal to the treatment effect for compliers. Alternative to the I.V. methods is the parametric and more restrictive *"Two Stages Model"*. It is essentially a simultaneous linear equations system, with a latent endogenous variable and a multivariate

TABLE 2. *T-test for the causal effect of the assignment on the treatment (df=6562).*

Effect	*t-test*	*p-value*
+0.0757	8.0642	0.000

TABLE 3. *Estimated causal effect of credit cards for the compliers (I.V. method).*

Effect	Stand.Dev.	*p-value*
+20.52	48.19	0.670

normal stochastic term:

$$\begin{cases} D_i^* = \alpha_{1H} + \mathbf{X}_i'\beta_{1XH} + \gamma_i \\ Y_i = \alpha_{2H} + \beta_{DH}D_i + \mathbf{X}_i'\beta_{2XH} + \varepsilon_i \,. \end{cases}$$

Where:

$$\begin{cases} D_i = 1 \text{ if } D_i^* \geq 0, \\ D_i = 0 \text{ if } D_i^* < 0; \end{cases}$$

$$cov\left(X_i, \gamma_i\right) = cov\left(X_i, \varepsilon_i\right) = 0;$$

$$\begin{pmatrix} \gamma_i \\ \varepsilon_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & \sigma^2 \end{pmatrix}\right];$$

$$\begin{pmatrix} \gamma_i \\ \varepsilon_i \end{pmatrix} \perp \begin{pmatrix} \gamma_j \\ \varepsilon_j \end{pmatrix} \qquad \forall\, i \neq j.$$

Under these and the further assumption of homogeneity (stating that causal treatment effect has to be equal for every unit), the A.T.E. for the whole population is identified and consistently estimated by a two-steps procedure (Heckman, 1978; 1979). In order to improve the fitting a set of pre-treatment variables ($\mathbf{X}$) is included in the two equations.

3 Results

In this application the role of random assignment to treatment can be attributed to the binary indicator of living, for a family, in close *"proximity*

TABLE 4. *Estimated causal effect of credit cards on Italian families' liquidity (in Euro); Two Stages Model.*

	parameter	*p-value*
Intercept	23.58	0.057
Effect of the use of credit cards	**-68.30**	**0.035**
Family size	+8.32	0.000
Geograph. area:		
Center Italy	-5.41	0.248
South Italy	+40.48	0.000
Town size (# inhabitants):		
beetwen 40.000 and 500.000	-2.01	0.769
beetwen 20.000 and 40.000	-11.31	0.133
less than 20.000	-8.48	0.280
Householder schooling:		
primary school	+7.21	0.392
junior high school	+20.71	0.020
high school	+42.81	0.000
University degree	+91.70	0.000
Householder position:		
white-collar	+13.19	0.081
manager, high officials	+23.81	0.175
professional man, entrepreneur	+43.40	0.000
unemployed	-17.17	0.173
housewife, retired	+30.53	0.000

to the bank". This choice is supported by the fact that the causal effect of the variable *"proximity to the bank"* on the treatment is significantly different from zero (Table 2). Despite the goodness of the instrument, the I.V. method produces an estimate of the effect of D_i on Y_i not significantly different from zero (Table 3). This result justifies the use of the parametric and more restrictive *"Two Stages Model"*.
The final result is a negative and significant causal effect of credit cards on the minimal amount of cash held by Italian families. The effect is quantified in -68.30 Euro for the families with at least a credit card holder (Table 4).

Acknowledgments: I would like to thank Fabrizia Mealli, Gliberto Ghilardi, an anonymous referee, as well as seminar participants in Firenze (I), and Pisa (I) for useful comments and suggestions.

References

Angrist, J.D., Imbens, G.W., and Rubin, D.B. (1996). Identification of causal effect using instrumental variables. *Journal of the American Statistical Association*, **91**, 444–455.

Banca d'Italia (1997). *Supplementi al Bollettino Statistico: I bilanci delle famiglie italiane nell'anno 1995*; Anno VII, n°14.

Heckman, J.J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica*, **46**, 931–959.

Heckman, J.J. (1979). Sample selection bias as a specification error. *Econometrica*, **47**, 153–161.

Holland, P.W (1986). Statistics and casual inference. *Journal of the American Statistical Association*, **81**, 945–960.

Imbens, G.W. and Angrist, J.D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, **62**, 467–475

Variance Free Model of Griffing's Type II Diallel Cross Experiments

Joao T. Mexia[1] and Stanisław Mejza[2]

[1] Departamento 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 diallel cross experiments in which the offsprings were obtained by Griffing's Type II crossing system. Moreover, an estimation and testing hypotheses concerning such genetical characteristics as general combining ability, specific combining ability and heterosis effect are discussed.

Keywords: Diallel cross system; General combining ability; Specific combining ability; Variance free model.

1 Introduction

In the breeding programs breeders wish to compare the performance of inbred lines, and in particular which crosses would be the most profitable. In this program two facts are very important, namely the crossing system of inbred lines (sometimes called matting design) and the diallel cross experiment (called environmental design). The statistical analysis deals with observations made on offspring obtained in some crossing system. In the paper we consider the system dealing with crossing the pairs of inbred lines. Two such systems are well known, i.e. *line x tester system* and *diallel crossing system.*

The statistical analysis of line x tester experiments by the approach of variance free model is given by Mejza and Mexia (2002).

By diallel crossing system we mean the one in which a set of p inbred lines is chosen and crosses among these lines are made. It means that we can get maximal pxp combinations. Diallel cross system may depending upon whether or not the parental inbred lines or the reciprocal F_1 are included or not. Griffing (1956) gave the classification of diallel cross systems. Additionally, he gave the statistical analysis for proposed four types of diallel crosses for data obtained in the experiments carried out in randomized complete block design. This analysis was generalized for other block design (cf. Mejza, 1994, Mejza and Mejza, 1995).

This paper deals with Griffing's II Type of crossing system.This type of diallel crossing includes parents and one set of $F_1^{'}$s (but not reciprocal $F_1^{'}$s).
In the selection process some genetical characteristics such as *general combining ability* (gca), *specific combining ability* (sca) and *heterosis effect* (he) play the crucial role (cf. Singh and Chandhary, 1979). The problem which naturally arises is the choice of proper environmental design. In the paper we assume that the diallel cross experiment was performed in a complete randomized design.

2 The Analysis of Diallel Crosses - Type II

Let us assume that the k-th replication y_{ijk} concerning the genotype obtained after crossing the i-th line with the j-th line, (shortly denoted as *(i,j)* - th cross) is modelled as follows:

$$y_{ijk} = \gamma_{ij} + e_{ijk} \quad i = 1, 2, ..., p,\ j = i, i+1, ..., p,\ k = 1, 2, ..., n, \qquad (1)$$

where γ_{ij} denotes the expected value of the trait observed on the cross *(i,j)* and e_{ijk} denotes the error. It will be assumed that $e_{ijk} \sim N(0, \sigma^2)$ for all *i, j, k*. The genotype effect can be expressed as:

$$\gamma_{ij} = \mu + g_i + g_j + s_{ij}, \qquad (2)$$

where μ denotes the general mean, $g_i, (g_j)$ - the gca effect of the i-th (j-th) line, s_{ij} - the sca effect of the *(i,j)* cross, such that $s_{ij} = s_{ji}$.
The gca, sca and he are defined by Griffing (1956) as follows:

$g_i = \frac{1}{(p+2)}(\gamma_{i.} + \gamma_{ii} - \frac{2}{p}\gamma_{..})$,

$s_{ij} = \gamma_{ij} - \frac{1}{(p+2)}(\gamma_{i.} + \gamma_{.j} + \gamma_{ii} + \gamma_{jj}) + \frac{2}{(p+1)(p+2)}\gamma_{..}$,

$h_{ij} = \gamma_{ij} - \frac{1}{2}(\gamma_{ii} + \gamma_{jj})$ or $h_{ij} = \gamma_{ij} - \max(\gamma_{ii}, \gamma_{jj})$,

where

$\gamma_{i.} = \sum_{j=i}^{p} \gamma_{ij}, \quad \gamma_{..} = \sum_{i=1}^{p} \sum_{j=i}^{p} \gamma_{ij}$,

¿From the definitions of gca and sca effects it follows that

$\sum_{i=1}^{p} g_i = 0$ and $\sum_{j=1}^{p} s_{ij} + s_{ii} = 0$, for each i.

Now let us define the general hypotheses that to be tested by the cross experiments considered.
The hypotheses can be expressed as:

1. H_{01}: $g_i = 0$; for all i,
2. H_{02}: $s_{ij} = 0$; for all $i \leq j$,
3. H_{03}: $g_i = 0$; for fixed i
4. H_{04}: $g_i - g_j = 0$; $i \neq j$,
5. H_{05}: $s_{ij} = 0$; for fixed i, j, $i \leq j$,
6. H_{06}: $s_{ij} - s_{ik} = 0$; $i \leq j, k, j \neq k$,
7. H_{07}: $s_{ij} - s_{kl} = 0$; $i \leq j, k \leq l, i \neq k, l, j \neq k, l$,
8. H_{08}: $h_{ij} = 0$; for all $i < j$,

where $i, j, k, l = 1, ..., p$. The above hypotheses can be verified by using standard analysis of variance technique.

3 Variance Free Model

Let us assume that on each offspring cross 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 offsprings from the mating design $(x_{ij1}, y_{ij1}), ..., (x_{ijn}, y_{ijn})$.
The inference concerning genotypes (genetical characteristics) can be based on these traits independently. But this is correct only when traits are uncorrelated (independently distributed). However, many times the traits are correlated and then is worth taking this fact into account in further inference from breeding experiments. In this paper we propose one of the ways allowing us to infer on genotypes (throughout the gca, sca and he) on the basis of correlation coefficients.
Let ρ_{ij} *i=1,2,..., p, j= i,i+1,...,p* be the correlation coefficient for the *(i,j)* cross and let r_{ij} be its estimator. Then using the transformation (cf. Kendal and Stuart, 1958, Mexia, 1990)

$$z_{ij} = 0.5\sqrt{n-3}\, ln((1 + r_{ij})/(1 - r_{ij})) \tag{3}$$

we obtain $z_{ij} \sim N(\mu_{ij}, 1)$, where

$$\mu_{ij} = 0.5\sqrt{n-3}\, ln((1 + \rho_{ij})/(1 - \rho_{ij})) + (\rho_{ij}\sqrt{n-3})/(2(n-1)),$$

$i = 1, 2, ..., p, \;\; j = i, i+1, ..., p.$
When the number of cross replication is quite large, the component $(\rho_{ij}\sqrt{n-3})/(2(n-1))$ is proportionally small with respect to the first part of μ_{ij}. Hence, in the further considerations we will assume that $z_{ij} \sim N(\tilde{\mu}_{ij}, 1)$, where $\tilde{\mu}_{ij} = 0.5\sqrt{n-3}\, ln((1 + \rho_{ij})/(1 - \rho_{ij})) = c\, ln\varphi_{ij}$,

where $c = 0.5\sqrt{n-3}$, $\varphi_{ij} = (1+\rho_{ij})/(1-\rho_{ij})$.

Finally, expressing $\tilde{\mu}_{ij}$ in the same way as in (2) we obtain the model

$$\tilde{\mu}_{ij} = \tilde{\mu} + \tilde{g}_i + \tilde{g}_j + \tilde{s}_{ij}, \tag{4}$$

where $\tilde{\mu}$ is the general mean, $\tilde{g}_i$, $\tilde{g}_j$ are the lines gca effects, respectively and $\tilde{s}_{ij}$ are the sca effects, $\tilde{s}_{ij} = \tilde{s}_{ji}$.
Hence, the z_{ij} we can express as:

$$z_{ij} = \tilde{\mu}_{ij} + \tilde{e}_{ij}, \tag{5}$$

where $\tilde{e}_{ij} \sim N(0,1)$.

Model (5) is called variance free model of Griffing's Type II cross experiment. To find the estimators of gca, sca and he in the model (5) we can use analysis of variance technique for two-factor experiments without replicates. For the model (5) the least squares estimators of gca, sca and he can be obtained by using z_{ij} instead of γ_{ij}. Let us note that the sums of squares in analysis of variance follow χ^2 distribution with known variance. This simplifies procedures for testing hypotheses.
The problem worth noticing is connected with the meaning of the hypotheses considered in the model (2) in relation to variance free model (5).
Let us introduce the following abbreviations:

$$\varphi_{i*} = \Pi_{j=1}^{p}\varphi_{ij}, \quad \varphi_{**} = \Pi_{u=1}^{p}\Pi_{v=u}^{p}\varphi_{uv}, \quad \kappa = (\varphi_{**})^{2/p}.$$

Then we have: $\tilde{\mu}_{i\cdot} = c\ ln\ \varphi_{i*}$, $\tilde{\mu}_{\cdot\cdot} = c\ ln\varphi_{**}$,
and
$g_i = (c/(p+2))\ ln((\varphi_{ii}\varphi_{i*})/\kappa)$,

$$s_{ij} = (c/(p+2))ln((\varphi_{ij}^{p+2}\varphi_{**}^{2/(p+1)})/(\varphi_{i*}\varphi_{j*}\varphi_{ii}\varphi_{jj})),$$

$h_{ij} = c\ ln\ (\varphi_{ij}/\sqrt{\varphi_{ii}\varphi_{jj}})$ or $h_{ij} = c\ ln\ (\varphi_{ij}/max(\varphi_{ii},\varphi_{jj}))$.

The hypotheses mentioned earlier can be expressed in the following way:

1. H_{01}^*: $\varphi_{ii}\varphi_{i*} = \kappa$; for all i.
2. H_{02}^*: $\varphi_{ij}^{p+2}\varphi_{**}^{2/(p+1)} = \varphi_{i*}\varphi_{j*}\varphi_{ii}\varphi_{jj}$; for all $i,j; i \le j$,
3. H_{03}^*: $\varphi_{i*}\varphi_{ii} = \kappa$; for fixed i,
4. H_{04}^*: $\varphi_{ii}\varphi_{i*} = \varphi_{jj}\varphi_{j*}$, $i \neq j$
5. H_{05}^*: $\varphi_{ij}^{p+2}\varphi_{**}^{2/(p+1)} = \varphi_{ii}\varphi_{i*}\varphi_{jj}\varphi_{*j}$; for fixed $i,j; i \le j$,
6. H_{06}^*: $\varphi_{ij}^{p+2}/(\varphi_{jj}\varphi_{*j}) = \varphi_{ik}^{p+2}/(\varphi_{kk}\varphi_{k*})$, $i \neq j$, k; $j \neq k$,

7. H_{07}^*: $\varphi_{ij}^{p+2}/(\varphi_{ii}\varphi_{i*}\varphi_{jj}\varphi_{j*}) = \varphi_{kl}^{p+2}/(\varphi_{kk}\varphi_{k*}\varphi_{ll}\varphi_{l*})$,
$i \leq j, k \leq l, i \neq k, l, j \neq k, l$,

8. H_{08}^*: $\varphi_{ij} = max(\varphi_{ii}, \varphi_{jj})$ or $\varphi_{ij} = \sqrt{\varphi_{ii}\varphi_{jj}}$,

where $i, j, k, l = 1, 2, ..., p$.

Acknowledgments: The paper was partially supported by KBN grant 6 P06A 026 21.

References

Griffing, B. (1956). Concept of general and specific combining ability in relation to diallel crossing systems. *Australian Journal of Biological Science*, **9**, 463–493.

Kendal, M. and Stuart, A. (1958). *The advanced Theory of Statistics* - Vol. I, Charles Griffin.

Mejza, I. and Mejza, S. (1995). An analysis of diallel progenies compared in a row-column design. *Journal of Genetics and Breeding*, **49**, 223–228.

Mejza, S. (1994). Diallel crossing systems in designs with orthogonal block structure. *Proc. 3rd Schwerin Conference on Mathematical Statistics, Selection Procedures II.*, FBN Dummerstorf, 32-36.

Mejza, S. and Mexia, J.T. (2002). Variance free model of line x tester experiments. *Statistical Modelling in Society. Proceedings of the 17th International Workshop on Statistical Modelling Chania, Crete. M. Stasinopoulos and G. Touloumi eds.* 453–458.

Mexia, J.T. (1990). Variance free models, *Trabalhos de Investigacao, Dept. of Mathematics, F.C.T., U.N.L.*, **2**.

Singh, R.K. and Chandhary, B.D. (1979). *Biometrical Methods in Quantitative Genetics Analysis.* New Delhi: Kallyani Duplisher.

The Consequence of Ignoring a Level of Nesting in Multilevel Analysis

Mirjam Moerbeek[1]

[1] Department of Methodology and Statistics, Utrecht University, PO Box 80140, 3508 TC Utrecht, The Netherlands, M.Moerbeek@fss.uu.nl

Abstract: Multilevel analysis is an appropriate tool for the analysis of hierarchically structured data. There may, however, be reasons to ignore one of the levels of nesting in the data analysis. In this paper we use a three level model with one predictor variable as a reference model and ignore the top or intermediate level in the data analysis. Analytical results show that this has an effect on the estimated variance components and that variances of estimated regression coefficients may be overestimated, leading to a lower power of the test of the effect of the predictor variable. These results depend on the ignored level and the level at which the predictor variable varies, as well as on the sample sizes and the variance components.

Keywords: Multilevel model; Variance components; Iterative generalized least squares; Power.

1 Introduction

In many studies data have a nested structure which means that persons are nested within clusters. Examples are pupils nested in classes nested in schools and employees nested within worksites. The correct statistical technique for the analysis of such data is multilevel analysis (e.g. Hox, 2002) since it accounts for correlated outcomes of persons within the same cluster.

Even when the multilevel model is used, incorrect conclusions may be drawn when not all possible levels for which variations in the outcome variable of interest occur are included into the model. There are several reasons for ignoring a level of nesting in a multilevel analysis. First, the data set does not include identifiers on all possible levels at which the outcome variable of interest varies. Second, applied researchers may found a model with many levels too complicated and may tend to use a simpler model. Third, in some cases certain levels of nesting are not clearly identified. For instance, groups of friends within schools are not as clearly identified as classes within schools. Fourth, levels of nesting may have to be ignored when the computer package to be used for the data analysis can only handle a limited number of levels.

The consequences of ignoring the top level in a two-level model have been extensively studied (Moerbeek, Van Breukelen and Berger, in press). Little is known on the consequence of ignoring a level of nesting in a multilevel model with more than two levels. Hutchison and Healy (2001) and Tranmer and Steel (2001) showed the effect of ignoring a level of nesting on the estimated variance components. Opdenakker and Van Damme (2000) analyzed an existing data set with four levels of nesting to show the effect of ignoring a level of nesting on the estimated parameters and their standard errors. The purpose of the present paper is to give a systematic overview of the consequence of ignoring a level of nesting in a multilevel analysis.

2 Multilevel Model

Assume the underlying data structure has three levels of nesting. For the sake of concreteness, units at the three levels are called pupils, classes and schools. For pupil i in class j in school k the multilevel model that relates outcome y_{ijk} to predictor variable x_{ijk} is given by

$$y_{ijk} = \gamma_0 + \gamma_1 x_{ijk} + v_k + u_{jk} + e_{ijk} \tag{1}$$

were $e_{ijk} \sim N(0, \sigma_e^2)$, $u_{jk} \sim N(0, \sigma_u^2), v_k \sim N(0, \sigma_v^2)$ are the random terms at the pupil, class and school level, respectively. In order to derive formulae that are of practical use we assume a balanced design with n_3 schools, n_2 classes per school, and n_1 pupils per class. The predictor variable may be measured at the pupil, class or school level. A predictor variable that is measured at the pupil level is assumed to be a pure pupil-level variable; that is, it varies at the class level only and its mean is the same in each class. Likewise, a variable measured at the class level is assumed to be a pure class level variable.

3 Effect of Ignoring a Level of Nesting on Estimated Variance Components

Iterative Generalized Least Squares (Goldstein, 1989) may be used for the analysis of model (1). Ignoring the class level results in

$$\begin{aligned} \widetilde{\widehat{\sigma}}_v^2 &= \widehat{\sigma}_v^2 + \frac{n_1 - 1}{n_1 n_2 - 1}\widehat{\sigma}_u^2 \\ \widetilde{\widehat{\sigma}}_e^2 &= \widehat{\sigma}_e^2 + \frac{n_1 n_2 - n_1}{n_1 n_2 - 1}\widehat{\sigma}_u^2 \end{aligned} \tag{2}$$

where the estimated variance components obtained with ignoring a level of nesting are indicated with a tilde in order to distinguish them from those for three levels of nesting. So, the variance component is redistributed over the other two variance components, depending on the sample sizes at the

TABLE 1. $\widehat{var}(\widehat{\gamma}_1)$ *without and with ignoring a level of nesting.*

level of variation	$\widehat{var}(\widehat{\gamma}_1)$ for three levels	ignored level	$\widehat{var}(\widehat{\gamma}_1)$ with ignoring the given level
pupil	$\frac{\widehat{\sigma}_e^2}{n_1 n_2 s_x^2}$	school	$\frac{\widehat{\sigma}_e^2}{n_1 n_2 n_3 s_x^2}$
pupil	$\frac{\widehat{\sigma}_e^2}{n_1 n_2 n_3 s_x^2}$	class	$\frac{\widehat{\sigma}_e^2 + \frac{n_1 n_2 - n_1}{n_1 n_2 - 1}\widehat{\sigma}_u^2}{n_1 n_2 n_3 s_x^2}$
class	$\frac{\widehat{\sigma}_e^2 + n_1\widehat{\sigma}_u^2}{n_1 n_2 n_3 s_x^2}$	school	$\frac{\widehat{\sigma}_e^2 + n_1(\widehat{\sigma}_u^2 + \widehat{\sigma}_v^2)}{n_1 n_2 n_3 s_x^2}$
school	$\frac{\widehat{\sigma}_e^2 + n_1\widehat{\sigma}_u^2 + n_1 n_2\widehat{\sigma}_v^2}{n_1 n_2 n_3 s_x^2}$	class	$\frac{\widehat{\sigma}_e^2 + n_1\widehat{\sigma}_u^2 + n_1 n_2\widehat{\sigma}_v^2}{n_1 n_2 n_3 s_x^2}$

class and pupil level. For any fixed n_1, the fraction of $\widehat{\sigma}_u^2$ that is added to $\widehat{\sigma}_e^2$ is equal to 0 if $n_2 = 1$, and increases if n_2 increases. For any fixed n_2, the fraction of $\widehat{\sigma}_u^2$ that is added to $\widehat{\sigma}_e^2$ is equal to 1 if $n_1 = 1$, and decreases if n_1 increases. Of course, the change in the estimated variance components at the pupil and school level is low if $\widehat{\sigma}_u^2$ is low. The situation is less complex for ignoring the school level, which results in

$$\begin{aligned} \widehat{\widehat{\sigma}}_u^2 &= \widehat{\sigma}_u^2 + \widehat{\sigma}_v^2 \\ \widehat{\widehat{\sigma}}_e^2 &= \widehat{\sigma}_e^2. \end{aligned} \tag{3}$$

So the estimated variance component at the school level is added to that at the class level, while that at the pupil level remains unchanged.

4 Effect of Ignoring a Level of Nesting on Test Statistics

Ignoring a level of nesting also has an effect on the $\widehat{var}(\widehat{\gamma}_1)$. For each of the three levels at which the predictor variable may vary Table 1 shows the $\widehat{var}(\widehat{\gamma}_1)$ for the three level model and for the model with ignoring the class or school level. s_x^2 is the variance of the predictor variable x_{ijk}. Note that we only consider the cases in which the ignored level is not the level at which the predictor variable varies, since it is realistic to assume that all identifiers at this level are well registered and available in the data set. The $\widehat{var}(\widehat{\gamma}_1)$ is too large if the predictor variable varies at the pupil level and the class level is ignored. The overestimation of $\widehat{var}(\widehat{\gamma}_1)$ increases when n_2 and/or $\widehat{\sigma}_u^2$ increase, or when n_1 decreases. The $\widehat{var}(\widehat{\gamma}_1)$ is also too large when the predictor variable varies at the class level and the school level is ignored. In this case the overestimation of $\widehat{var}(\widehat{\gamma}_1)$ increases when n_1 and/or $\widehat{\sigma}_v^2$ increase, but does not depend on n_2. Moreover, Table 1 shows that ignoring the school level does not have an effect on the $\widehat{var}(\widehat{\gamma}_1)$ of a pupil level predictor, and that ignoring the class level does not have an effect on the $\widehat{var}(\widehat{\gamma}_1)$ of a school level predictor.

5 Conclusions and Discussion

This paper has shown that ignoring a level of nesting has an effect on the estimated variance components. The variance of a regression coefficient may be overestimated, leading to a too small test statistic for the test on the effect of the associated predictor variable on the outcome, and consequently too low power.
For simplicity a model with just one predictor variable was used in this paper, but it can be shown that the results also hold for models with more predictor variables as long as each predictor variable varies at just one level of the multilevel data structure. A predictor variable that varies at more that one level may be split up into orthogonal components that each vary at just one level (Neuhaus and Kalbfleisch, 1998). Of course the results only hold if the same predictors are used in the model with three levels of nesting and in the model with ignoring a level of nesting. The situation becomes more complex if predictor variables at the ignored level are also removed from the multilevel model since then the variance components, and hence standard errors of regression coefficients estimators, are not only affected by ignoring the level of nesting but also by the removal of these predictor variables from the multilevel model (Snijders and Bosker, 1994).
It should be noted that we assumed balanced designs since that leads to relatively simple formulae for the change in variance components when ignoring a level of nesting and to relatively simple formulae for the $\widehat{var}(\widehat{\gamma}_1)$. If sample sizes vary, their mean values may be substituted into these formulae, which then only hold approximately. The approximation is, of course, less accurate if the variability in sample sizes is large.

References

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, **73**, 43–56.

Hox, J.J. (2002). *Multilevel Analysis. Techniques and Applications*. New Jersey: Erlbaum.

Hutchison, D. and Healy, M. (2001). The effect of variance component estimates of ignoring a level in a multilevel model. *Multilevel Modelling Newsletter*, **13**, 4–5.

Moerbeek, M., Van Breukelen, G.J.P., and Berger, M.P.F. (in press). A comparison between traditional methods and multilevel regression for the analysis of multi-center intervention studies. *Journal of Clinical Epidemiology*.

Neuhaus, J.M. and Kalbfleisch, J.D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, **54**, 638–645.

Opdenakker, M.-C. and Van Damme, J. (2000). The importance of identifying levels in multilevel analysis: An illustration of the effects of ignoring the top or intermediate levels in school effectiveness research. *School Effectiveness and School Improvement*, **11**, 103–130.

Snijders, T.A.B. and Bosker, R.J. (1999). Modeled variance in two-level models. *Sociological Methods and Research*, **22**, 342–363.

Tranmer, M. and Steel, D.G. (2001). Ignoring a level in a multilevel model: Evidence from UK census data. *Environment and Planning A*, **33**, 941–948.

The Use of Score Tests for Inference on Variance Components

Geert Molenberghs[1] and Geert Verbeke[2]

[1] Limburgs Universitair Centrum, Center for Statistics, Universitaire Campus, Building D, B-3590 Diepenbeek, Belgium; geert.molenberghs@luc.ac.be
[2] Katholieke Universiteit Leuven, Biostatistical Center, UZ Sint-Rafaël, Kapucijnenvoer 35, B-3000 Leuven, Belgium; geert.verbeke@med.kuleuven.ac.be

Abstract: Whenever inference for variance components is required, the choice between one-sided and two-sided tests is crucial. This choice is usually driven by whether or not negative variance components are permitted. For two-sided tests, classical inferential procedures can be followed, based on likelihood ratios, score statistics, or Wald statistics. For one-sided tests, however, one-sided test statistics need to be developed, and their null distribution derived. While this has received considerable attention in the context of the likelihood ratio test, there appears to be much confusion about the related problem for the score test.

Keywords: Boundary condition; Likelihood ratio test; Linear mixed model; One-sided test; Variance component.

1 Introduction

The linear mixed-effects model (Laird and Ware 1982, Verbeke and Molenberghs 2000) is a commonly used tool for variance component models and for longitudinal data. Let $\boldsymbol{Y_i}$ denote the n_i-dimensional vector of measurements available for subject $i = 1, \ldots, N$. A general linear mixed model then assumes that $\boldsymbol{Y_i}$ satisfies

$$\boldsymbol{Y_i} = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i} + \boldsymbol{\varepsilon_i}, \tag{1}$$

in which $\boldsymbol{\beta}$ is a vector of population-averaged regression coefficients called fixed effects, and where $\boldsymbol{b_i}$ is a vector of subject-specific regression coefficients. The $\boldsymbol{b_i}$ describe how the evolution of the ith subject deviates from the average evolution in the population. The matrices X_i and Z_i are $(n_i \times p)$ and $(n_i \times q)$ matrices of known covariates. The random effects $\boldsymbol{b_i}$ and residual components $\boldsymbol{\varepsilon_i}$ are assumed to be independent with distributions $N(\mathbf{0}, D)$, and $N(\mathbf{0}, \Sigma_i)$, respectively. Inference for linear mixed models is usually based on maximum likelihood or REML *under the marginal model*. Thus, we can adopt two *different* views on the linear mixed model. The

fully *hierarchical* model is specified by

$$\begin{array}{rcl} \boldsymbol{Y_i}|\boldsymbol{b}_i & \sim & N_{n_i}(X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i, \Sigma_i), \\ \boldsymbol{b}_i & \sim & N(0, D), \end{array} \tag{2}$$

while the marginal model is given by

$$\boldsymbol{Y_i} \sim N_{n_i}(X_i\boldsymbol{\beta}, V_i = Z_i D Z_i' + \Sigma_i). \tag{3}$$

Even though they are often treated as equivalent, there are important differences between both views. Obviously, (2) requires the covariance matrices Σ_i and D to be positive definite, while in (3) it is sufficient for the resulting matrix V_i to be positive definite. Different hierarchical models can produce the same marginal model and some marginal models are not implied by any hierarchical model.
The simplest example to illustrate differences between the marginal and hierarchical views is found by restricting the random effects in (1) to a random intercept, producing the marginal model:

$$\boldsymbol{Y}_i \sim N(X_i\boldsymbol{\beta}, \tau^2 J_{n_i} + \sigma^2 I_{n_i}) \tag{4}$$

where J_{n_i} equals the $n_i \times n_i$ matrix containing only ones. In the marginal view, negative values for τ^2 are perfectly acceptable (Nelder 1954, Verbeke and Molenberghs 2000, Sec. 5.6.2), since this merely corresponds to negative within-cluster correlation $\rho = \tau^2/(\tau^2 + \sigma^2)$. In the hierarchical view, it is clearly imperative to restrict τ^2 to nonnegative values.

2 Inference for Variance Components

While each of the two views are possible, there are important differences regarding statistical inference for variance components. The first, *unconstrained case*, is classical regarding inference for the variance component τ^2 since the usual two-sided alternative $H_0 : \tau^2 = 0$ versus $H_{A2} : \tau^2 \neq 0$ is then used. Wald, likelihood ratio, and score tests are then asymptotically equivalent, and the asymptotic null distribution is well known to be χ^2_1. In the *constrained case*, one typically needs one-sided tests of the null-hypothesis

$$H_0 : \tau^2 = 0 \qquad \textit{versus} \qquad H_{A1} : \tau^2 > 0. \tag{5}$$

As the null-hypothesis is now on the boundary of the parameter space, classical inference no longer holds, appropriate tailored test statistics need to be developed, and the corresponding (asymptotic) null distributions derived. We will briefly review the likelihood-ratio case and then turn to score tests in the next section.
Suppressing dependence on the other parameters, let $\ell(\tau^2)$ denote the log-likelihood, as a function of the random-intercepts variance τ^2. Further, let

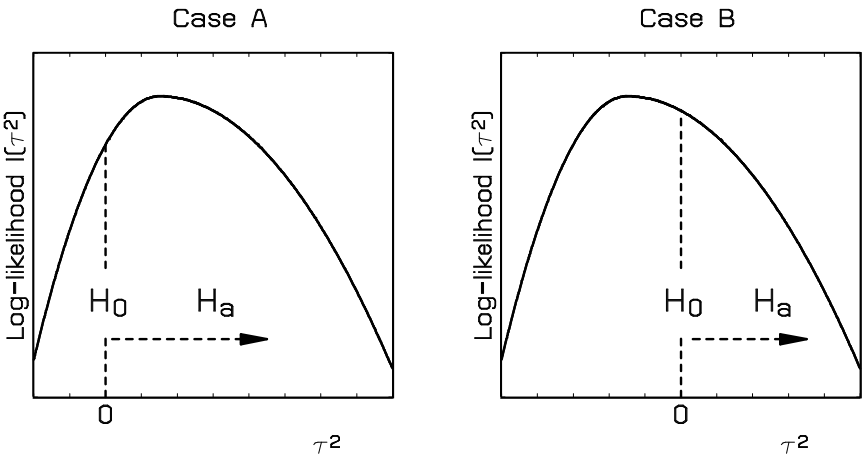


FIGURE 1. *Graphical representation of two different situations, when developing one-sided tests for the variance τ^2 of the random intercepts b_i in model.*

$\widehat{\tau}^2$ denote the maximum likelihood estimate of τ^2 under the unconstrained parameterization. We first consider the likelihood ratio test, with statistic:

$$T_{LR} \quad = \quad 2 \ln \left[\frac{\max_{H_{1A}} \ell(\tau^2)}{\max_{H_0} \ell(\tau^2)} \right].$$

Two cases, graphically represented in Figure 1, can now be distinguished. Under Case A, $\widehat{\tau}^2$ is positive, and the likelihood ratio test statistic is identical to the one that would be obtained under the unconstrained parameter space for τ^2. Hence, conditionally on $\widehat{\tau}^2 \geq 0$, T_{LR} has asymptotic null distribution equal to the classical χ^2_1. Under Case B, $\ell(\tau^2)$ is maximized at $\tau^2 = 0$ under H_{1A} as well as under H_0, yielding $T_{LR} = 0$. Both cases are equally probable to occur, under the null. Hence, the asymptotic null distribution of T_{LR} is easily seen to follow a $0.5P(\chi^2_1 > c) + 0.5P(\chi^2_0 > c)$ null distribution. This was one of Stram and Lee's (1994) special cases. Note that, whenever $\widehat{\tau}^2 \geq 0$, the observed likelihood ratio test statistic is equal to the one under the unconstrained model, but the p-value is half the size of the one obtained from the classical χ^2_1 approximation to the null distribution.

In general, inference under the unconstrained model for the variance components in D can be based on the classical chi-squared approximation to the null distribution for the likelihood ratio test statistic. Under the constrained model, Stram and Lee (1994) have shown that the asymptotic null distribution for the likelihood ratio test statistic for testing a null hypothesis which allows for k correlated random effects versus an alternative of $k+1$ correlated random effects (with positive semi-definite covariance ma-

trix D_{k+1}), is a mixture of a χ^2_k and a χ^2_{k+1}, with equal probability 1/2. For more general settings, e.g., comparing models with k and $k+k'$ ($k'>1$) random effects, the null distribution is a mixture of χ^2 random variables (Shapiro 1988), the weights of which can only be calculated analytically in a number of special cases. Shapiro's (1988) results provide a few important special cases, not studied by Stram and Lee (1994). For example, if the null hypothesis allows for k *un*correlated random effects (with a diagonal covariance matrix D_k) versus the alternative of $k+k'$ uncorrelated random effects (with diagonal covariance matrix $D_{k+k'}$), the null distribution is a mixture of the form

$$\sum_{m=0}^{k'} 2^{-k'} \binom{k'}{m} \chi^2_m.$$

Shapiro (1988) shows that, for a broad number of cases, determining the mixture's weights is a complex and perhaps numerical task.

3 The Score Test

Verbeke and Molenberghs (2003), using results by Silvapulle and Silvapulle (1995), have shown that similar results are obtained when a score test is used instead of a likelihood ratio test. The use of score tests for testing variance components under a constrained parameterization requires replacing the classical score test statistic by an appropriate one-sided version. This is where the general theory of Silvapulle and Silvapulle (1995) on one-sided score tests proves very useful. They consider models parameterized through a vector $\boldsymbol{\theta} = (\boldsymbol{\lambda}', \boldsymbol{\psi}')'$, where testing a general hypothesis of the form $H_0 : \boldsymbol{\psi} = \mathbf{0}$ *versus* $H_A : \boldsymbol{\psi} \in \mathcal{C}$ is of interest. Silvapulle and Silvapulle (1995) allow $\mathcal{C}$ to be a closed and convex cone in Euclidean space, with vertex at the origin. The advantage of such a general definition is that one-sided, two-sided, and combinations of one-sided and two-sided hypotheses are included.

Adopt the following notation. Let $\boldsymbol{S}_N(\boldsymbol{\theta})$ and $H\boldsymbol{\theta})$ be the score vector and Hessian matrix of the log-likelihood function. Further, decompose $\boldsymbol{S}_N$ as $\boldsymbol{S}_N = (\boldsymbol{S}'_{N\lambda}, \boldsymbol{S}'_{N\psi})'$, let $H_{\lambda\lambda}(\boldsymbol{\theta})$, $H_{\lambda\psi}(\boldsymbol{\theta})$ and $H_{\psi\psi}(\boldsymbol{\theta})$ be the corresponding blocks in $H(\boldsymbol{\theta})$, and define $\boldsymbol{\theta}_H = (\boldsymbol{\lambda}', \mathbf{0}')'$. $\boldsymbol{\theta}_H$ can be estimated by $\widehat{\boldsymbol{\theta}}_H = (\widehat{\boldsymbol{\lambda}}', \mathbf{0}')'$, in which $\widehat{\boldsymbol{\lambda}}$ is the maximum likelihood estimate of $\boldsymbol{\lambda}$, under H_0. Finally, let $\boldsymbol{Z}_N$ be equal to $\boldsymbol{Z}_N = N^{-1/2}\boldsymbol{S}_{N\psi}(\widehat{\boldsymbol{\theta}}_H)$. A one-sided score statistic can now be defined as

$$T_S := \boldsymbol{Z}'_N H^{-1}_{\psi\psi}(\widehat{\boldsymbol{\theta}}_H)\boldsymbol{Z}_N \;-\; \inf\left\{(\boldsymbol{Z}_N - \boldsymbol{b})' H^{-1}_{\psi\psi}(\widehat{\boldsymbol{\theta}}_H)(\boldsymbol{Z}_N - \boldsymbol{b}) | \boldsymbol{b} \in \mathcal{C}\right\}. \quad (6)$$

Note that the score statistic, heuristically defined in the case of the random-intercepts model is a special case of (6). Indeed, when $\widehat{\tau}^2$ is positive, the score at zero is positive, and therefore in $\mathcal{C}$, such that the infimum in (6)

becomes zero. For $\widehat{\tau}^2$ negative, the score at zero is negative as well and the infimum in (6) is attained for $\boldsymbol{b} = \mathbf{0}$, resulting in $T_S = 0$.
It follows from Silvapulle and Silvapulle (1995) that, under suitable regularity conditions, for $N \to \infty$, the likelihood ratio and score test statistics satisfy $T_{LR} = T_S + o_p(1)$. This indicates that the equivalence of the score and likelihood ratio tests not only holds in the two-sided but also in the one-sided cases. Moreover, what is known about the null distribution in the case of the likelihood ratio test, immediately carries over to the score test case. This result corrects the common belief that, even when variance components are on the boundary of the parameter space, the score test deserved no special treatment. Verbeke and Molenberghs (2003) provide an empirical illustration. In practice, calculation of (6) requires some extra programming work and, even though it is not insurmountable, in most situations one may therefore be inclined to resort to likelihood ratio testing.

Acknowledgments: We acknowledge support from FWO-Vlaanderen Research Project "Sensitivity Analysis for Incomplete and Coarse Data" and Belgian IUAP/PAI network "Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data".

References

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

Nelder, J.A. (1954). The interpretation of negative components of variance. *Biometrika*, **41**, 544–548.

Shapiro, A. (1988). Towards a unified theory of inequality constrained testing in multivariate analysis. *International Statistical Review*, **56**, 49–62.

Silvapulle, M.J. and Silvapulle, P. (1995). A score test against one-sided alternatives. *Journal of the American Statistical Association*, **90**, 342–349.

Stram, D.O. and Lee, J.W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, **50**, 1171–1177 (correction in 1995).

Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data*, Springer Series in Statistics, Springer-Verlag, New-York.

Verbeke, G. and Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics*, **59**, 254–262.

An Efficient Method to Estimate Multiple Mean-Shift Models

Vito M. R. Muggeo[1]

[1] GRASPA Research Group (PRIN MIUR 2000 *'Statistics in Environmental Risk Evaluation'*– grant MM13208412_001). Email: vito.muggeo@giustizia.it

Abstract: An efficient way to estimate mean-shift models is illustrated: the method is particulary attractive as it does not depend on the sample size and on the number of changepoints to be estimated. Explanatory variables are allowed.

Keywords: Changepoint; Mean-shift model; Non-monotonic trend; Nile data.

1 Introduction

Given a sequence of n time-ordered random variables $Y_1, \ldots, Y_n$, 'mean-shift models' are defined as models where the mean values change at some unknown time-points, the so-called *changepoints*. For instance with one changepoint ψ the model is $E[Y_t] = \beta$ for $t \leq \psi$ and $E[Y_t] = \beta + \beta_1$ for $t > \psi$. These have to be contrasted with models where the mean values are connected at the unknown *breakpoints*, hereafter referred as 'breakpoint models' or 'continuous changepoint models'. The theoretical, different, issues related to both the models, as well as practical difficulties in estimating them, are well-known. A common feature is that standard asymptotic theory is unreliable, the basic concept being that the statistic tests are not random variables but stochastic processes. As regards to hypothesis testing, the papers previously published on the topic may be divided into two wide categories: *i*) methods based on the celebrated 'CuSum approach' avoiding of estimating the model; *ii*) likelihood-ratio-type tests based on comparisons between the model with and without changepoint. The latter approach requires ML estimates to be available. With regard to estimation, most of works proposed in the literature use grid-search-type algorithms and thus have the disadvantage to depend strongly on the sample size, n, and the number of changepoints, L. A 'simple' grid-search algorithm works in $O(n^L)$ operations, but recently Bai and Parron (2003) discussed an improved 'dynamic' version working in $O(n^2)$ operations for any L. However with a modest series length and more than one changepoint, the dynamic programming can still be rather time-consuming and so this might be an obstacle in practice.

Here we illustrate a simple but very efficient method to estimate mean-shift models with any n and L. The method relies on an exact algorithm

recently proposed to estimate continuous changepoints in regression models (Muggeo, 2002). We illustrate the idea in the following sections.

2 Estimating the Model

Let $x = 1, 2, \ldots, n$ the time variable; omitting the indices and the error terms, a multiple ($L > 1$) continuous changepoint model for the response z is given by

$$z = \beta_0 + \beta x + \beta_1 (x - \psi_1) I(x > \psi_1) + \ldots + \beta_L (x - \psi_L) I(x > \psi_L) \quad (1)$$

This model implies gradual changes at the unknown breakpoints ψ_l ($l = 1, 2, \ldots, L$), where the generic difference-in-slope parameter, β_l, has to be non-null if the lth change occurs. The breakpoint model (1) is continuous at $x = \psi_l$ where its first derivative $z' = y$, namely

$$y = \beta + \beta_1 I(x > \psi_1) + \ldots + \beta_L I(x > \psi_L) \quad (2)$$

follows a multiple mean-shift model: here even the same regression function is discontinuous at the changepoints. Therefore there exists a direct correspondence between the equations (1) and (2): the latter is the first derivative with respect to x of the former, and the slope parameters in one are the intercepts in the other. Thus from a mean-shift model (2) for the observed y, we define the correspondent breakpoint model for z and estimate it by means of the exact algorithm discussed in Muggeo (2002): the method relies just on fitting iteratively a certain linear model and needs starting values for the changepoints that may be easily obtained by the, possibly smoothed, plot. A simple idea is used to build the 'working response' z: given $y_1, y_2, ..., y_n$ ordered according to the linear model $a + bx$ ($x = 1, \ldots, n$), say, it is well known that the differentiated values $\nabla y_t = y_t - y_{t-1}$ represent the first derivative of the ys with respect to x: the fitted line throughout the points $(x_t, \nabla y_t)$ is a null-slope-line with intercept equal to b, i.e. $E[\nabla y_t] = b$. Thus by the inverse argument, given observations arranged according to a parallel line $y = b$, say, it is possible calculate the relevant 'integrated' data lined with slope b:

1. fix $z_1 = k$ 2. calculate $z_t = y_{t-1} + z_{t-1}$ for $t = 2, \ldots, n + 1$

Hereafter we use Δy to mean integrated data obtained by the step 1. and 2. above: hence in short $\Delta y = z$ and $\nabla z = y$. If a mean-shift model (2) holds for y, the correspondent working response z follows a breakpoint model (1) with respect to x; the intercept depends on the $z_1 = k$, but this is not of interest, as it is not included in the model (2). Figure 1 displays some simulated mean-shift models and the corresponding breakpoint models.
The estimates $\hat{\beta}$, $\hat{\beta}_l$ and $\hat{\psi}_l$ for the breakpoint model are assumed as point estimates for the mean-shift model, therefore fitted values and residual dispersion may be easily obtained. Although the output of the breakpoint

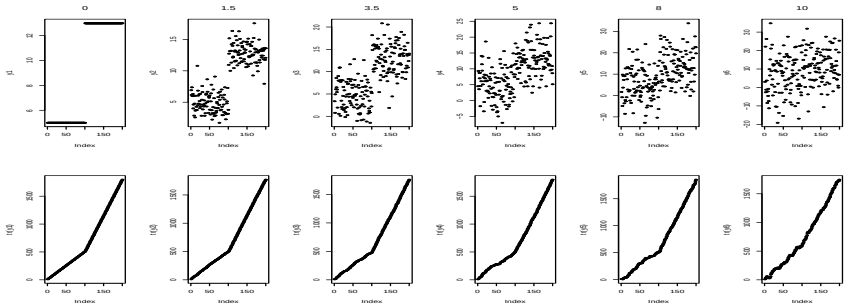


FIGURE 1. *Simulated mean-shift models* $y = 5 + 8 \times I(x > 100) + \epsilon$, $\epsilon \sim N(0, \sigma)$ *and relevant transformations; from left to right* $\sigma = 0, 1.5, 3.5, 5, 8, 10$. *The variance-reduction in breakpoint models is noticeable with* $\sigma = 10$.

model includes the $2(L+1)$ full covariance matrix, it does not seem possible to evaluate the $2L+1$ covariance matrix for the mean-shift model. In order to get standard errors and confidence intervals for the parameter estimates some resampling-based method could be employed, e.g. the bootstrap. It is worth noting that bootstrap methods are rather prohibitive in changepoint analysis since the estimation of the changepoints needs a lot of time for each model using grid-search-type algorithms, but the linear method (Li) here discussed does not suffer from such limitation. To study the performance of such method and compare it with the dynamic grid-search approach (Dy), a small simulation study was undertaken, generating $n = 100$ gaussian variates according to two different changepoint models ($L = 1$ and $L = 2$). Both the Li and Dy methods were applied and results are shown in Table 1.

TABLE 1. *Comparison between the 'dynamic' (*Dy*) and 'linear' (*Li*) algorithms: mean (m), median (m) and standard deviation (s) of the changepoint estimators (1000 replicates of* $n = 100$ *simulated gaussian data).*

	$L=1$			$L=2$					
	$\psi = 20$			$\psi_1 = 35$			$\psi_2 = 70$		
Method	m	*m*	s	m	*m*	s	m	*m*	s
Dy	25.3	20.0	13.3	35.5	35.0	7.4	69.1	70.0	7.1
Li	27.1	21.0	16.4	36.6	35.0	10.7	66.9	69.0	12.7

Results are, in general, according to the expectations. Performance of the estimators, in terms of bias and standard deviation, is poorer when the changepoint is on the edges (i.e. $\psi \to 0$ or n) and gets better as ψ moves near the middle ($\psi \to n/2$); however, even with $\psi = 35$ the estimators are substantially unbiased and for $\psi = 50$ they are so exactly (results not shown). Moreover both bias and standard deviation decrease as n and/or β_l increase. As one could expect it, Dy outperforms Li because it scans all

possible values, therefore bias and standard deviation of the DY-estimator are lower. But differences in the standard deviations are rather noticeable and although further research is needed, at least two remarks might be addressed: because the likelihood for ψ could not be strictly concave (as it is the case for linear parameters) it could happen that sometimes LI finds a local optima and stops; DY does not estimate the model simultaneously (it estimates ψ by grid-search and estimate the linear beta parameters *assuming fixed* $\hat{\psi}$), thus uncertainty of the beta-estimates might be ruled out in the psi-estimates. These reasons could cause the DY estimators to be less variable. Execution time was, of course, much lower in LI method; for $L = 1$ differences are trifling, but for $L = 2$ we recorded an average time to estimate one model by DY equal to 7.8 and 27.9 seconds for $n = 100$ and $n = 200$ respectively; LI used always less than 1.5 seconds, allowing resampling techniques to be employed in reasonable times. Calculations were performed using the R packages strucchange by A. Zeileis and the forthcoming segmented, both working with no external code.
Additional covariates with fixed coefficients can be included in the model in a straightforward way. In order to fit a mean-shift model with explanatory variables w, namely $y = \beta + \sum \beta_l I(x > \psi_l) + \gamma w$, it is sufficient to fit $\Delta y = \beta_0(k) + \beta x + \sum \beta_l (x - \psi_l) I(x > \psi_l) + \gamma \Delta w$, namely the integrated transformations have to be applied at every explanatory variable w: this is the analogous in differentiated model where the effect of w on y is the same of one of ∇w on ∇y.

3 Application: the Nile Data-set

Figure 2 illustrates the time series of the Nile whose annual flows seem to drop in 1898 because a dam was built; this is a well-known data-set in changepoint problems, see for instance Cobb (1978). According to the method discussed above, a continuous changepoint model is applied to the working response. Even if the plot shows a possible break-point located at $x = 30$ approximatively, the starting value is set 60 in order to show that the starting points are, generally, a minor issue. Figure 2 shows the observed data along with the fitted values and the 'transformed data' on which a breakpoint model is estimated. The estimated changepoint is the 28^{th} observation, namely the 1898 as elsewhere reported. The estimated slopes before and after the break-point for the transformed 'z-data' are 1082.37 and 851.01 respectively, corresponding to the mean levels in the original data-set.

4 Conclusions

An efficient algorithm to estimate mean-shift models has been illustrated. Emphasis is given on the estimation problem that seem to have received

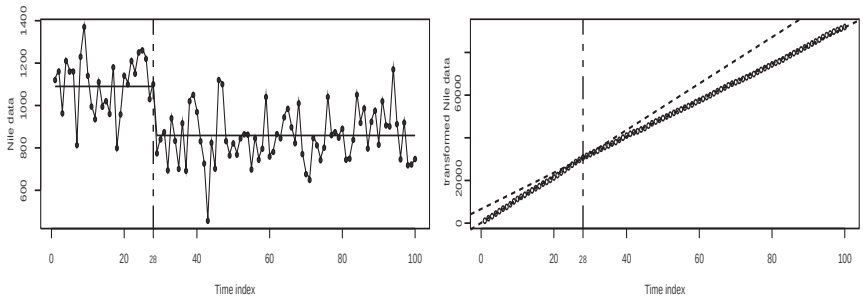


FIGURE 2. *The Nile data-set: observed and fitted values (left) and transformed 'z' working values (right); broken lines emphasizes the breakpoint at* $x = 28$.

less attention in literature. Of course given the point estimates, one could apply likelihood-based tests that are known to have greater power when the change is expected to take place on the tails.

Although it has to be acknowledged that the method has to be investigated further to assess, for instance, effect of autocorrelation on the transformed data and possible bias in the estimates of additional explanatory variables, the method is quite attractive and seems to work well in practice; Its main features include: i) exactness, i.e. if the breakpoints exist they are always revealed in data with relatively low variance; ii) noticeable efficiency, being the estimation substantially independent of n and L; iii) ability in estimating simultaneously the model, and therefore allowing (co)variability among the estimates to be taken correctly into account. Furthermore it should be noted that the transformed 'integrated' data are much less scattered around the straight lines. Therefore passing from a mean-shift to a breakpoint model, allows to deal with much more clear-cut relationships making changepoint detection and estimation easier. As example see the last picture on the right in Figure 1, where $\sigma = 10$: the two parallel lines in the mean shift model are almost negligible, but the breakpoint is rather evident through the z-values. Of course when the variance is very high relatively to parameter values (in the example in Figure 1, $\sigma = 50$), the derived V-shaped relationship can be very wiggly and breakpoint detection is still difficult.

References

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, **18**, 1–22.

Muggeo, V.M.R. (2002). Estimating regression models with unknown breakpoints. *Statistics in Medicine*. (in press).

Cobb, G.W. (1978). The problem of the Nile: Conditional solution to a change-point problem. *Biometrika*, **65**, 243–251.

Correcting for Inter-observer Variability in a Geographical Oral Health Study

Samuel M. Mwalili[1], Emmanuel Lesaffre[1], and Dominique Declerck[2]

[1] Biostatistical Centre, Katholieke Universiteit Leuven
Kapucynenvoer 35, B-300 Leuven, Belgium.
samuel.mwalili@med.kuleuven.ac.be
[2] School of Dentistry, Katholieke Universiteit Leuven, Belgium

Abstract: Scoring of caries experience in dental studies usually involves different examiners. In this respect, the kappa statistic (Cohen, J. 1960) is often used to indicate the inter-examiner variability. We argue that the kappa values are not useful to indicate the impact of the examiner effect on an epidemiological study. Instead, we present a logistic random effects model with a correction term. The correction term for inter-examiner bias with respect to a gold standard is obtained from a calibration study. As an example the proposed approach was applied to a geographical oral health study based on the Signal Tandmobiel® dataset. A frequentist as well as a Bayesian approach were used with the latter providing an easy way of accounting for the variability of the estimated regression coefficients. Further, we evaluated how large the calibration data should be to obtain reliable estimates.

Keywords: Calibration study; Error-in-variables; Inter-examiner variability; Ordinal logistic regression.

1 Introduction

From a dental point of view it was of interest to examine the geographical trend in caries experience in Flanders. To this end, we employed the Signal Tandmobiel® data, a 6 year longitudinal oral health study started in Flanders (Belgium) in 1996 involving 4468 children. The outcome of interest is the *dmft* index, which is the sum of the number of decayed (d), missing due to caries (m) and filled (f) teeth. The *dmft* index is hereby referred to as the caries experience. To examine the geographical trend first a random logistic regression model was applied (Hartzel, Agresti and Caffo (2001)) whereby

$$log\left(\frac{\pi_{ik1}+\ldots+\pi_{ikr}}{\pi_{ik,r+1}+\ldots+\pi_{ik4}}\right)=\lambda_r+\boldsymbol{x}_i'\boldsymbol{\beta}+u_k, \qquad r=1,2,3 \qquad (1)$$

where $\boldsymbol{x_i}$ is a d-dimensional vector of covariates pertaining to the ith child and $\boldsymbol{\beta}$ is the corresponding vector of regression coefficients (fixed effects), π_{ikr} is the probability of child i in school k being classified in category r of the ordinal caries response. The random intercept u_k pertains to the kth school and we assume that $u_k \sim N(0, \sigma^2)$. λ_1 is the intercept and (λ_2, λ_3) are the ordered category cut-off parameters of the *dmft* index, which satisfy $\lambda_2 < \lambda_3$. Below the vector $(\lambda_1, \lambda_2, \lambda_3)'$ will be denoted as $\boldsymbol{\lambda}$. Observe that only the caries experience of the first year of examination was used. In this preliminary analaysis both frequentist (SAS procedure NLMIXED) and Bayesian (WINBUGS program) models were fitted.

In a Bayesian context the likelihood needs to be combined with a prior distribution of the parameters. First, we choose the same prior distribution for the regression coefficient β_s $(s = 1, \cdots, d)$, i.e, $\beta_s \sim N(0, 10^{-6})$. The precision is chosen to be very small so that a vague prior distribution for $\boldsymbol{\beta}$ is obtained, ensuring that the posterior is data driven. Second, we choose $\sigma^2 \sim IG(10^{-2}, 10^{-2})$ so that the prior distribution is sufficiently diffuse. Third, we take a vague normal prior for λ_1, i.e., $\lambda_1 \sim N(0, 10^{-6})$ and a truncated normal prior for the other category cutoffs, i.e., $\lambda_2 \sim N(0, 10^{-6})I(\lambda_1, \lambda_3)$ and $\lambda_3 \sim N(0, 10^{-6})I(\lambda_2, +\infty)$.

The results of these fit clearly indicated a significant East-West gradient in the degree of caries experience, being higher in the East of Flanders. However, this model did not take into the account the scoring variability of the 16 dental examiners, each active in a restricted geographical area.

A calibration dataset contrasting the sixteen dental examiners involved in the study against a gold standard on scoring *dmft* was available. Hence, we could calculate the kappa statistics for each examiner against this gold standard. From a scheme proposed by Landis and Koch (1977) we concluded that our kappa values showed moderate to good agreement. However, it is not directly clear what message these kappa values brings us on the outcome of the dental analysis.

A classical way to take a confounder into account is to include it into the (logistic) regression model. Clearly, controlling for examiner removed the geographical East-West trend. The same conclusion could be drawn when the examiner was included in the model as a random effect. We argue that correcting for examiner in this way is not appropriate because it does not take into account the scoring bias and/or variability of the examiners. Instead, we opted for another correction.

2 Methodology

To properly take the examiners' effect into account, we propose an ordinal random effect logistic regression model with a *correction term*. This correction term could be estimated from the calibration data. Our model is

given by,

$$\Pr(y_{ij} \leq a|\boldsymbol{\gamma}, \boldsymbol{x_i}, u_k) = \sum_{c=1}^{a}\sum_{d=1}^{4}\gamma_{jcd}q_{ikd}, \tag{2}$$

whereby

$$q_{ik} = \begin{pmatrix} F(\lambda_1 + \boldsymbol{x_i}'\boldsymbol{\beta} + u_k) \\ F(\lambda_2 + \boldsymbol{x_i}'\boldsymbol{\beta} + u_k) - F(\lambda_1 + \boldsymbol{x_i}'\boldsymbol{\beta} + u_k) \\ F(\lambda_3 + \boldsymbol{x_i}'\boldsymbol{\beta} + u_k) - F(\lambda_2 + \boldsymbol{x_i}'\boldsymbol{\beta} + u_k) \\ 1 \qquad\qquad - F(\lambda_3 + \boldsymbol{x_i}'\boldsymbol{\beta} + u_k) \end{pmatrix}.$$

$q'_{ik} = (q_{ik1}, \cdots, q_{ik4})$ and γ_{jab} is the conditional probability of classifying a discretized *dmft* score in the ath category by examiner j given it is classified in the bth category by the gold standard.

The SAS procedure NLMIXED (SAS Institute Inc.) can be used to estimate the unknown parameters ($\boldsymbol{\lambda}$, $\boldsymbol{\beta}$, σ^2) of model (2) when for $\boldsymbol{\gamma}$ an estimated value is imputed from the calibration data. However, one needs to take into the account the uncertainty with which the $\boldsymbol{\gamma}$ is estimated. This could be done by some analytical approximations but it can be even done more easily by a Bayesian approach, e.g. by using WINBUGS. We have followed this approach to analyse our geographical dental study.

Finally, we provide analytical as well as simulation results to illustrate the impact of calibration sample size on the precision of the corrected model. Firstly, we have taken examiners with five different scoring behaviours: (1) severely underscoring, (2) moderately underscoring, (3) variable (even in underscoring & overscoring), (4) moderately overscoring, and (5) severely overscoring all with the same kappa value compared with the gold standard. Secondly, we used five examiners. Each examiner was biased in scoring caries experience as compared to the gold standard and was active in only one of the five provinces of Flanders. Thus our second simulation study mimicked the epidemiological dental study.

3 Results

The Bayesian posterior estimates of geographical regression coefficients from WINBUGS applied to the Signal Tandmobiel® study are slightly larger in absolute value compared to the estimated regression coefficients from NLMIXED, but overall the estimates of the NLMIXED and WINBUGS are close, given their estimated variability. Further, a sensitivity analysis by varying different prior distributions showed that our conclusions were relatively stable. Compared to the NLMIXED output, the standard errors of our estimated Bayesian regression coefficients are higher, due to the variability with which $\boldsymbol{\gamma}$ is estimated from the calibration data. But more importantly, the East-West gradient remained important in both geographical models.

The results of the first simulation exercise are displayed in Table 1. These results show that a kappa value of less than 1 is always associated with attenuated estimated regression coefficients. This is known from the error-in-variables literature Carroll *et al* (1995). It is also known that the size of calibration dataset (n_0) has a great impact on the correction for bias and the (increased) variability of the corrected estimates.

TABLE 1. *Simulation study 1: Parameter estimates from a logistic binary regression model where the response is measured with error. The examiner has* $\kappa = 0.6$ *compared to the gold standard with five different scoring patterns.*

		Gold		corrected		
Pattern	Parameter	Standard mean(sd)	Crude mean(sd)	$n_0 = 50$ mean(sd)	$n_0 = 100$ mean(sd)	$n_0 = 200$ mean(sd)
1	$\beta_0 = 0$	-0.01(.09)	-0.52(.09)	-0.03(.37)	-0.01(.30)	0.01(.24)
	$\beta_1 = 1$	1.02(.14)	0.58(.13)	1.16(.69)	1.12(.61)	1.07(.36)
2	$\beta_0 = 0$	0.00(.09)	-0.30(.09)	-0.05(.45)	0.01(.31)	-0.01(.25)
	$\beta_1 = 1$	1.00(.13)	0.56(.13)	1.16(.82)	1.08(.43)	1.05(.39)
3	$\beta_0 = 0$	0.00(.09)	-0.00(.09)	-0.03(.41)	-0.01(.32)	-0.04(.25)
	$\beta_1 = 1$	1.00(.14)	0.57(.12)	1.17(.80)	1.08(.40)	1.05(.30)
4	$\beta_0 = 0$	0.00(.09)	0.30(.09)	-0.04(.42)	-0.03(.32)	0.01(.24)
	$\beta_1 = 1$	1.00(.13)	0.61(.13)	1.12(.65)	1.05(.40)	1.03(.28)
5	$\beta_0 = 0$	-0.01(.09)	0.51(.09)	-0.06(.49)	-0.02(.31)	-0.06(.24)
	$\beta_1 = 1$	1.01(.14)	0.66(.14)	1.13(.63)	1.05(.29)	1.02(.25)

The results of the second simulation study are shown in Table 2. Now two covariates were involved covariate 1 (effect β_1) and a geographical east-west covariate (effect β_2).

TABLE 2. *Simulation study 2: Parameter estimates from a logistic binary regression model where the response is measured with error controlling for covariate 1 and a geographical covariate. Each examiner has* $\kappa = 0.6$ *against the gold standard.*

Parameter	Gold Standard mean(sd)	Crude mean(sd)	corrected $n_0 = 50$ mean(sd)	corrected $n_0 = 100$ mean(sd)	corrected $n_0 = 200$ mean(sd)
$\beta_0 = 0$	-0.005(.08)	0.974(.08)	0.054(0.56)	0.013(0.38)	0.006(.27)
$\beta_1 = 1$	0.997(.08)	0.402(.06)	0.921(0.28)	0.959(0.22)	0.978(.19)
$\beta_2 = 2$	2.017(.14)	-0.729(.12)	1.893(1.35)	1.995(0.98)	2.002(.66)

By correcting for misclassification on the response, it is clear that the correction reduces bias in the estimates to 0, but with an increase in the

standard errors of the estimates. The bias and standard errors of the estimates from the corrected model decreases as the number of observations in the calibration data increases. In application to the Signal Tandmobiel® data, this model model confirmed the East-West gradient in the degree of caries experience in Flanders.

4 Conclusion

The simulation studies show that the correction terms removes the bias introduced by the misclassification of the response by the examiner with respect to the gold standard. However, the correction reduces the precision of the parameter estimates. These results are equally supported by estimates of the ordinal logistic applied to the Signal Tandmobiel® study. We find that the larger the calibration sample size the larger the precision and reduction of the bias after correction.

References

Carroll, R.J, Ruppert, D., and Stefanski, L.A. (1995). *Non-linear Measurement Error Models*. London: Chapman & Hall.

Cohen J. (1960). A coefficient of agreement for nominal scales.*Educational and Psychological Measurement*, Vol. **XX**(1).

Gelman, A., Carlin, B.J., Stern, S.H., and Rubin, B. D. (1995). *Bayesian Data Analysis*. London: Chapman & Hall.

Hartzel, J., Agresti, A., and Caffo, B. (2001). Multinomial logit random effects models. *Statistical Modelling*, **1**, 81–102.

Landis, J.R. and Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

Mwalili, S.M., Lesaffre, E.M., and Declerck, D. (2003). A bayesian correction for an ordinal response measurement error in a geographical oral health study. Submitted.

Richardson, S. and Gilks, W.R. (1993). A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *American Journal of Epidemiology*, **138**, 430–432.

SAS Institute Inc. (1999-2001). *The SAS System for Windows*. Cary, NC, USA.

Spiegelhalter, D., Thomas, A., Best, N., and Gilks, W. (1996). *Bayesian Inference Using Gibbs Sampling Manual (version ii)*. Cambridge, UK.

Extended Likelihood Inference applied to a New Class of Models

John A. Nelder[1]

[1] Department of Mathematics, Imperial College, 180 Queen's Gate, London SW7 2AZ, UK. Email: j.nelder@imperial.ac.uk

1 From GLMs to HGLMs

A GLM has the following three-part structure:
(i) A response vector y with mean μ and independent errors from a one-parameter exponential family.
(ii) A set of covariates $X_1, X_2, ..., X_k$, whose effects β combine linearly to form a linear predictor,

$$\eta = \sum X_i \beta_i.$$

(iii) A monotonic link function $\eta = g(\mu)$ connecting the linear predictor and the mean.
The classical normal model has errors from a normal distribution and link function the identity function, $\eta = \mu$. A GLM may have error distributions including Poisson, binomial, multinomial, gamma and inverse Gaussian. Any GLM distribution may be expressed by its variance, which has the form

$$var(y) = \phi V(\mu),$$

where ϕ is the dispersion parameter and $V()$ is the variance function. The kernel of the log-likelihood has the form

$$\sum \{y\theta - b(\theta)\}/\phi,$$

where $\theta = \theta(\mu)$ is the GLM canonical parameter. If $\eta = \theta$, the link is a canonical link. GLMs have a single algorithm for fitting any model of the class. It generalizes the least-squares of classical models to iterative weighted least squares, using an adjusted dependent variate

$$z = \eta + (y - \mu)(\partial\eta/\partial\mu)$$

in place of y, with iterative weights given by

$$W = (\partial\mu/\partial\eta)^2 V(\mu)^{-1}.$$

Goodness of fit is measured by the deviance, a log-likelihood-ratio statistic which generalizes the residual sum of squares.
HGLMs extend GLMs in two important ways. First the mean and the dispersion may be modelled jointly, and secondly the linear predictor may contain both fixed and random effects, each with its own dispersion parameter to be estimated. These two extensions are discussed below.

2 Joint Modelling of Mean and Dispersion

Such models can be expressed as two interlinked GLMs, one for the mean and one for the dispersion. Given values of ϕ, the dispersion parameter (now varying over experimental units), the reciprocals are used as prior weights for the analysis of the mean as a GLM. From the analysis of the mean, we form the deviance components for each unit, and these become the response for the dispersion GLM. The distribution for the dispersion GLM is gamma, and the link is usually assumed to be the log. A linear predictor on the log scale is postulated, and the joint fit is done by alternating mean and dispersion models until convergence. For details see Lee & Nelder (1998). An important field of application for this technique is that of quality-improvement experiments.

3 Random Effects

Random effects are well known in the normal case, the linear predictor $X\beta$ being extended to $X\beta + Zu$, where u is a vector of random effects from a normal distribution. Estimates are required for the fixed effects and the variance components for y and u. Individual estimates of the random effects u are called best unbiased linear predictors or BLUPs. In HGLMs the response y may follow a GLM and the distribution of the random effects may come from any conjugate distribution of a GLM. Such conjugate distributions include the gamma, inverse gamma, beta and normal.

4 Fitting an HGLM

Two criteria are used for fitting an HGLM. For the fixed and random effects beta and u, given the dispersion components, we maximize the hierarchical or h-likelihood, first defined by Lee & Nelder (1996). This consists of two parts, one derived from the conditional distribution of $y|u$, and one from the distribution of the random effects. The random effect u may appear in the linear predictor for y on some scale $v(u)$, say, $v = \log(u)$, in which case the second term is derived from the density of v not u. The h-likelihood is not a Fisherian likelihood because the random effects are not observed, but we believe it to be the natural extension of Fisher likelihood to random-effect models. For given values of the dispersion components, we fit the

fixed effects β and the random effects u simultaneously, by maximizing the h-likelihood. To estimate the dispersion components we use a generalization of REML, using an adjusted profile h-likelihood eliminating the fixed and random effects. We alternate between estimation of β and u given the dispersion components, and of the dispersion components given β and u.

5 The Algorithm

The fitting of a HGLM can be reduced to the fitting of interlinked GLMs, as in the joint modelling of mean and dispersion. The model for the mean can be reduced to an augmented GLM, in which the response vector is augmented by quasi-data for the random effects, and the augmented design matrix (for one random effect) has the form $\begin{pmatrix} X & Z \\ 0 & I \end{pmatrix}$. The model for the dispersion can be written as a GLM with response derived from the deviance component for the model for the mean. Again, the fitting is done alternately until convergence. See Lee and Nelder (2001a) for details.

6 Extensions

The model class can be expanded, first to allow correlated observations expressed by random effects in the linear predictor (Lee and Nelder,2001b), and secondly to allow random effects in the dispersion model. The latter gives rise to Double HGLMs, and these have important implications in expanding the class of financial models. For a fuller list of the applications of HGLMs see Lee (2003).

7 Software

Software for fitting HGLMs and DHGLMs is available as a set of Genstat procedures: the procedures include those for setting up the model, fitting the model, display of results and model checking. The reduction of the fitting to that of interlinked GLMs implies that GLM model-checking techniques can be extended to the wider class,

References

Lee, Y. (2003). Hierarchical generalized linear 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. 257–261.

Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B*, **58**, 619–678.

Lee, Y. and Nelder, J.A. (1998). Generalized linear models for the analysis of quality-improvement experiments. *Canadian Journal of Statistical Society, Series B*, **58**, 619–678.

Lee, Y. and Nelder, J.A. (2001a). HGLMs: A synthesis of GLMs, random effect models and structured dispersions. *Biometrika*, **88**, 987–1006.

Lee, Y. and Nelder, J.A. (2001b). Modelling and analysing correlated non-normal data. *Statistical Modeling*, **1**, 7–16.

Screening for Outliers From a Log-linear Model

Gerhard Neubauer[1] and Johannes Hofrichter[1]

[1] Institute of Applied Statistics, Joanneum Research, Steyrergasse 25a, 8010 Graz, Austria. Email: gerhard.neubauer@joanneum.at

Abstract: Identification of unbalanced regional supply is one approach to financial controlling of health care systems. Hospital discharge data together with demographic data make a simple supply indicator. Analysing such data from Styria by a main effect Poisson model shows very bad model fit. 80% of the observations are identified as outliers and the interpretation of these results would lead to assess the local hospital system absolute supply deficiency. Improving the fit by modelling overdispersion through the quasi likelihood approach (QLM) and the negative binomial model (NBM) is successful. Only of few observations remain outliers and consequently the supply with hospital care is judged satisfying. As the QLM and the NBM identify different sets of outliers a bootstrap approach to variance estimation is applied. It turns out that the same observations are identified as with the QLM.

Keywords: Residuals; Overdispersion; Bootstrap.

1 Introduction

This paper was motivated by analyses of data from the Styrian health care system. As a consequence of growing concern about the costs of the public health care, financial controlling based on empirical data became important. One approach to controlling intends to optimize the allocation of money by detecting regional over- and undersupply with medical services. Hospital discharge frequency is used as supply indicator and the regional distribution of the population serves as benchmark. Assuming equal hospitalization rate in all regions supply deficiency is indicated by large deviations of the hospital discharge distribution from the (benchmark) population distribution.

We use hospital discharge data from the six Styrian NUTS-3 (NUTS=Nomenclature des unités territoriales statistiques) regions that contain ICD-9 (ICD=International Classification of Diseases) diagnostic information, so that 17 main disease groups can be distinguished. Thus we have a total of $6 \times 17 = 102$ observations Y_{ij} ($i = 1, \ldots, 17$ disease groups DG_i, and $j = 1, \ldots, 6$ regions R_j). The research problem is to discover regional differences

in hospitalization frequency for the 17 disease groups which can be stated as testing

> H_0: The rate of hospitalization due to DG_i is the same for all regions, i.e. $\lambda_{i1} =, \ldots, = \lambda_{iJ} = \lambda_i$.

versus

> H_1: The rate of hospitalization due to DG_i differs for at least two regions, i.e. $\lambda_{ij} \neq \lambda_{ik}$ for some pair (j, k).

Let $Y_{ij} \sim P(\mu_{ij})$ then under H_0 we have

$$\lambda_i = \mu_{ij}/n_j \quad \text{or} \quad \mu_{ij} = n_j\lambda_i,$$

and under H_1

$$\lambda_{ij} = \mu_{ij}/n_j \quad \text{or} \quad \mu_{ij} = n_j\lambda_{ij},$$

with n_j the population in R_j. Testing H_0 vs. H_1 is an ANOVA-type of problem for rates. Using the normal approximations to the Poisson variables the statistics

$$T_i = \sum_j \frac{(Y_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \tag{1}$$

are χ^2-distributed with $df = 5$ under H_0. Taking $\hat{\lambda}_i = n_i/n$ as estimate we obtain $\hat{\mu}_{ij} = (n_i n_j)/n$, where $n_i = \sum_j Y_{ij}$ and $n = \sum_j n_j$. In cases with $T_i > \chi^2_\alpha$ H_0 is rejected and at least one region has a deviant hospitalization rate.

2 Log-linear Models

2.1 Poisson Regression

The problem described in section 1 can be represented by a GLM (McCullagh & Nelder, 1989), namely the Poisson regression using the canonical log-link. Under H_0 we have

$$\mu^{(0)}_{ij} = n_j\lambda_i = n_j \exp(\beta_i), \tag{2}$$

which can be seen as a series of 17 intercept models with offset n_j, and (1) is just the Pearson chi-square statistic X^2_i. To test for regional differences the H_0-models are estimated and X^2_i and the deviances D_i should have values smaller than the critical value $\chi^2_\alpha(df)$. The estimation of the 17 models is equal to the simultaneous estimation of 17 intercepts within one model. Consequently $X^2 = \sum_i X^2_i$ and $D = \sum_i D_i$. Thus equation (2) can also be seen as a main effect Poisson model (PM) for the hospitalization rates with factor "Disease group".

2.2 Interaction, Bias and Outlier Detection

For H_1 we have the interaction model

$$\mu_{ij}^{(1)} = n_j \lambda_{ij} = n_j \exp(\beta_i + \gamma_{ij}) = \mu_{ij}^{(0)} \exp(\gamma_{ij}). \tag{3}$$

Model (3) is not applicable to our data as it is the saturated model with $\hat{\mu}_{ij}^{(1)} = Y_{ij}$. But the residuals from model (2) can be used to obtain some information about the unknown γ_{ij}. Based on the estimates from model (2) we obtain $E(r_{ij}^{(0)}) = E(Y_{ij} - \hat{\mu}_{ij}^{(0)}) = 0$ given that H_0 is valid. If H_1 holds then the estimate $\hat{\mu}_{ij}^{(0)}$ is biased and

$$E(r_{ij}^{(1)}) = E(Y_{ij} - \hat{\mu}_{ij}^{(0)}) = \mu_{ij}^{(1)} - \mu_{ij}^{(0)} = \mu_{ij}^{(0)}[\exp(\gamma_{ij}) - 1] \neq 0.$$

This bias carries over to the standardized Pearson residuals

$$sr_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}^{(0)}}{[(1 - h_{ij})V(\hat{\mu}_{ij}^{(0)})]^{1/2}}, \tag{4}$$

which are asymptotically $N(0,1)$ under H_0 ($V(\cdot)$, the variance function, h_{ij}, the diagonal elements of the GLM hat matrix). Therefore $|sr_{ij}| > z_{1-\alpha/2}$ can be used as screening tool to identify large deviations from the means, i.e. biased observations. In regression analysis observations with values of sr_{ij} outside of some interval (e.g. $[-1.96, 1.96]$) are regarded as outliers. Estimating model (2) when actually H_1 is correct results in biased estimates for μ_{ij} and consequently bad model fit. Besides misspecification of the mean (leading to bias) we have misspecification of the variance function (leading to wrong precision) as second important reason for bad model fit. Bias enters the statistics X^2 and sr_{ij} in the numerator, while the variance assumption affects the denominator. Thus observing bad model fit together with many outliers may be caused by heavy bias and/or the wrong variance model.

2.3 Modelling Dispersion

As we cannot estimate a better mean model to achieve better model fit, we look for alternatives to the Poisson variance assumption $V(\mu) = \mu$. In general $V(\cdot)$ may depend on some additional scale or dispersion parameters ϕ or α, like for instance the QLM with $V(\mu, \phi) = \phi\mu$, and the NBM with $V(\mu, \alpha) = \mu + \alpha\mu^2$. For $\phi > 1$ and $\alpha > 0$ we model overdispersion, i.e. $V(\cdot) > \mu$. A consequence of using $V(\cdot) > \mu$ is that only large deviations are identified as outliers. We expect to find smaller sets of biased observations for the sr_{ij} from the QLM and the NBM. Besides that the sets may differ markedly as $\phi\mu > \mu + \alpha\mu^2$ for $0 < \mu < \phi/\alpha$, and vice versa for $\mu > \phi/\alpha$.
ϕ is usually estimated by X^2/df or D/df, while α can be estimated by maximum likelihood. For the QLM the scaled value of X^2 is just df when

$\hat{\phi} = X^2/df$ is used, and $df(X^2/D)$ when $\hat{\phi} = D/df$ is used (similar for D). Thus the improvement in model fit for the QLM cannot be properly expressed in a number like for the NBM. This leaves us with the problem to choose between the QLM and the NBM.

3 Bootstrap

In section 2 we use the asymptotic properties of the standardized Pearson residuals to identify outliers. The results may depend strongly on the assumed variance structure and moreover there is no ad hoc method to choose between the competing models QLM and NBM. As alternative for outlier identification the nonparametric bootstrap shall be used to estimate the distribution of the residuals. The bootstrap method based on residual resampling for GLM (Davidson & Hinkley, 1999) requires that the residuals are approximately iid, which is best met by the standardized deviance residuals

$$rd_{ij} = \frac{sign(y_{ij} - \hat{\mu}_{ij})[2(l(y_{ij}) - l(\hat{\mu}_{ij})]^{1/2}}{(1 - h_{ij})^{1/2}} = \frac{d(y_{ij}, \hat{\mu}_{ij})}{(1 - h_{ij})^{1/2}}.$$

To obtain the bootstrap response y^*_{ij} we sample from the set of centered rd_{ij} ($e^*_{ij} \in \{rd_{ij} - \overline{rd}_{ij}\}$) and solve the equation $e^*_{ij} = d(y^*_{ij}, \hat{\mu}_{ij})$ by the Newton method. Fitting the bootstrap models $\mu^*_{ij} = n_j \exp(\beta^*_i)$ we estimate the variance of the raw residuals $r_{ij} = y_{ij} - \hat{\mu}_{ij}$ by

$$Var^*(r_{ij}) = \frac{1}{B-1} \sum_{k}^{B} (y^{*(k)}_{ij} - \hat{\mu}^{*(k)}_{ij})^2, \tag{5}$$

where $(\cdot)^{*(k)}$ $(k = 1, \ldots, B)$ are the bootstrap replications. To identify the outliers we compare the residuals standardized with (5) given by

$$sbr_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{Var^*(r_{ij})^{1/2}},$$

with the percentiles of the empirical distribution of all standardized bootstrap residuals given by

$$sbr^*_{ij} = \frac{y^*_{ij} - \hat{\mu}^*_{ij}}{Var^*(r_{ij})^{1/2}}.$$

4 Application and Results

Application of the 17 intercept Poisson models from equation (2) to our data shows that H_0 has to be rejected for all disease groups at both usual critical values $\chi^2_{0.05}(5) = 11.07$ and $\chi^2_{0.01}(5) = 15.86$ (see *Table 1*).

TABLE 1. *Results of the Poisson models for the 17 disease groups* DE_i

DG_i	X_i^2	D_i	DG_i	X_i^2	D_i
01	219.72	214.16	10	423.32	411.88
02	616.63	600.36	11	93.93	95.87
03	83.36	84.91	12	343.30	330.28
04	40.61	42.27	13	1499.12	1440.24
05	1322.94	1305.58	14	17.60	17.36
06	1748.23	1699.09	15	322.87	313.25
07	1191.09	1150.52	16	1282.92	1188.75
08	224.31	221.08	17	1366.61	1357.18
09	301.42	301.27			

TABLE 2. *Results for the Poisson Model, the Quasi-Likelihood Model and the Negative Binomial Model*

					Number of outliers	
	X^2	D	$\hat{\phi}$	$\hat{\alpha}$	$\alpha = 0.05$	$\alpha = 0.01$
PM	11097.97	10774.06	–	–	81	78
QLM	11097.97	10774.06	126.75	–	8	4
NBM	107.73	103.55	–	0.0232	8	3

For the PM we also have a very bad model fit: $X^2 = 11097.97$ and $D = 10774.06$ with $df = 85$. This is paralleled by 80% of the observations identified as outliers, i.e. $|sr_{ij}| > z_{1-\alpha/2}$, and further $|sr_{ij}| < 33$ with mean (standard deviation) 0.79 (11.63). The interpretation of these results is simply that the Styrian hospital system is doing extremely bad in supplying the Styrian population with health care.

Fitting the QLM and the NBM has the expected effect. The standardized Pearson residuals are now much smaller for both the QLM ($|sr_{ij}| < 3$) and the NBM ($|sr_{ij}| < 5$). The mean (standard deviation) is now 0.07 (1.03) for the QLM, and -0.001 (1.13) for the NBM. The number of identified outliers is reduced drastically for the QLM and the NBM (see *Table 2*), and consequently the supply of the Styrian population with health care from hospitals can be judged as balanced, save for a few exceptions. Table 2 gives some results for the PM, QLM and NBM.

A closer look on the outliers shows that the QLM and the NBM identify different observations. At $\alpha = 0.05$ four out of twelve observations are identified by both models, and at $\alpha = 0.01$ the models only agree on one out of six observations. This divergence is due to the different shapes of the estimated variance functions (cf. section 2.3).

The bootstrap method based on $B = 1999$ samples gives estimates of $Var^*(r_{ij})$ that lead to bootstrap standardized residuals $|sbr_{ij}| < 3$ with mean (standard deviation) 0.05 (1.13). The empirical distribution of the standardized bootstrap residuals sbr^*_{ij} has mean (standard deviation) 0.00 (1.00), but it is slightly skewed and shows a peak. Using the $\alpha/2$ and $1-\alpha/2$ quantiles of this distribution we identify the same outliers as the QLM. A possible explanation for this result is given by Friedl (1997), who showed that resampling from standardized Pearson residuals yields a bootstrap dispersion estimate that is just the empirical variance of the Pearson residuals. Moreover this estimate can be used as an estimate for the dispersion parameter ϕ of the QLM. It appears that this also holds approximately for the standardized deviance residuals.
For final interpretation of the results we combine the two sets of outliers by union, giving twelve suspicious observations, and intersection, which gives us four substantial outliers ($\alpha = 0.05$). The variance functions of the QLM and the NBM cross at $\mu = 0$ and $\mu = \phi/\alpha$. The intersection is equivalent to using the upper part of both functions, i.e. $V(\mu) = \phi\mu$ for $0 < \mu < \phi/\alpha$ and $V(\mu) = \mu + \alpha\mu^2$ for $\mu \geq \phi/\alpha$, and the union takes the lower parts.

Acknowledgments: We are grateful to Herwig Friedl for helpful discussions and comments on our work.

References

Davison A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and their Applications.* Cambridge: CUP.

Friedl, H. (1997). On the asymptotic moments of Pearson type statistics based on resampling procedures. *Computational Statistics*, **12**, 265–277.

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

Bias of Logits in Environmental Impact Studies

Sandra Nunes[1], Joo Tiago Mexia[2], and Christoph Minder[3]

[1] Departamento de Matemática, Escola Superior de Tecnologia de Setúbal, Instituto Politécnico de Setúbal, Campus do IPS, Estefanilha, 2914-508 Setúbal, Portugal
[2] Departamento de Matemática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Quinta da Torre, 2829-516 Caparica, Portugal
[3] Department of Social and Preventive Medicine, University of Berne, Finkenhubelweg 11, 3012 Berne, Switzerland

Abstract: In environmental epidemiology, as in any other subject involving observational studies, one is interested in finding effect estimates that are (almost) unbiased and have smallest possible variance. In environmental epidemiology, as in other fields, one can achieve this only through a proper design. Logit models are a precious toll in this area.

Keywords: Logit models; Environmental impacts; Bias; Measurement errors.

1 Introduction

Logit models are widely used to express the probability of occurrences of diseases as a function of environmental impacts.
Typically, exposures are measured at fixed stations. These exposures are then taken to be representative for the whole population near this station. As a consequence, both measurement errors and measurements biases must be taken into consideration.
In this presentation, we consider a situation with measurement error only. We use such a model for discussing what happens when the error in the measurement of the impacts is not negligible. Our main result is that the bias in the slope of the response is always negative. A result like this one expresses a loss of sensibility to the exposures variations that results from less precision in their measurement.

2 Model

We assume that with exposure f the probability of disease is given by

$$p = \frac{e^{\beta_0+\beta_1 f}}{1 + e^{\beta_0+\beta_1 f}}.$$

Let $\hat{\beta}_1$ and $\tilde{\beta}_1$ be the maximum likelihood estimators obtained from exact $f_1, \ldots, f_n$ measures of exposure and approximate measures $\tilde{f}_1, \ldots, \tilde{f}_n$ with

$$\tilde{f}_i = f_i + \varepsilon_i, i = 1, \ldots, n$$

the $\varepsilon_1, \ldots, \varepsilon_n$ being independent and identically distributed with null mean value and variance σ_ε^2.
If we have n samples with dimensions n_i , $i = 1, \ldots, n$, the corresponding logits having variances v_i , $i = 1, \ldots, n$, it may be shown through lengthy but straightforward computations that

$$E\left(\tilde{\beta}_1 - \hat{\beta}_1\right) = -\beta_1 \frac{\left[\left(\sum_{i=1}^n \frac{1}{v_i}\right)^2 - \sum_{i=1}^n \frac{1}{v_i^2}\right]}{\left(\sum_{i=1}^n \frac{1}{v_i}\right)\left(\sum_{i=1}^n \frac{f_i^2}{v_i}\right) - \left(\sum_{i=1}^n \frac{f_i}{v_i}\right)^2} \sigma_\varepsilon^2 < 0.$$

We point out that the denominator in this expression may be used to estimate the variance component σ_f^2 , associated to exposures differences between the chosen stations.
The next step was to partition the variance σ_ε^2 in two independent components

$$\sigma_\varepsilon^2 = \sigma_a^2 + \sigma_e^2$$

where

- σ_a^2 – is the variance component associated to the sampling errors inside the population associated to the monitoring stations;
- σ_e^2 – is the variance component associated to measurement errors.

Through more straightforward computations it may be shown that

$$E\left(\tilde{\beta}_1 - \hat{\beta}_1\right) = -\beta_1 \frac{\sigma_a^2 + \sigma_e^2}{\sigma_f^2}.$$

3 Scenarios

We now apply the previous expression to study measurement design in several scenarios.

3.1 First Scenario

The monitoring stations as well as the sub-population are chosen. For each station there will be a corresponding region, so the station as well as the region will be given. Of the three variance components we can only influence σ_e^2 . Let us assume that there is a relation between costs and precision for this component, given by a decreasing function with an horizontal asymptote.
Our aim is to determinate the point C' of cost per station to the right of which there is only limited decrease in σ_e^2.

3.2 Second Scenario

The monitoring stations are given but the regions limits are not. Let us assume that the population assigned to a station is formed by L sub-populations. Since we have n populations we will have the same number of decompositions. Besides this for each station there will be a sub-population, for which it is most representative, the standard sub-population.
In this scenario we can influence (minimize) both σ_a^2 and σ_e^2 , subject to the use of certain sub-populations, thus controlling σ_ε^2 and through it the bias.
A rule that we can follow, is to assign each sub-population to the station for which the standard population is nearest to it. To carry out this work we need to choose:

- the standard sub-population for each station;
- a dissimilarity measure between sub-populations;
- to build the dissimilarity matrix between sub-populations.

3.3 Third Scenario

The third scenario is a free scenario. Let us assume that we only have an approximate idea of the number of stations to implant. We could start from the dissimilarity matrix that we previously considered and apply a cluster analysis to try to identify sub-populations that are gravity centers. These could be the chosen as standard sub-populations for the monitoring stations. It could happen that not all of the proposed stations could be implanted. It is important to point out that in this last scenario we can influence the three components σ_a^2 , σ_e^2 and σ_f^2 . Thus achieving a better control of the bias than in the first two scenarios.
Our aim is to minimize

$$\frac{\sigma_a^2 + \sigma_e^2}{\sigma_f^2}.$$

If we assume that we are comparing scenarios for which the values of $\sigma_a^2+\sigma_e^2$ are similar, we then could think of maximizing σ_f^2.

4 The Future Work

To work on Scenarios 2 and 3 we must have information to enable us to obtain the above mentioned dissimilarity matrix.

4.1 Required Information

In order to collect such an information we should start with a list of variables that influence exposure and characterize populations. The ideal would be:

1. To select the variables in the population;
2. To define the sub-population;
3. To build the dissimilarity matrix;
4. To select the sub-population as candidates to be standard sub-population;
5. To select the standard populations and to define the "limits (boundaries)";
6. To implant the monitoring stations;
7. To build a sample in each region.

At the end we should, using the approximated (adjusted) values, obtain the average incidence $\hat{p}$.
On the other hand the chosen variables to define the population should include residence and/or working place in order to facilitate establishing the regions boundaries.
Let us observe that there is a certainty likeness between the method that we are trying to develop and the "a priory tariffation method. Thus we can start with a preliminary large list of discrete variables, and then, check which of those are significantly related to incidence variations. The next step should be the use of the chosen variables to define sub-populations.

References

Cochran, W.G. (1977). *Sampling Techniques.* 3rd edition. New York: Wiley.

Finney, D.J. (1971). *Probit Analysis.* 3rd edition. Cambridge University Press.

Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression.* 2nd edition. Wiley Series in Probability and Statistics.

Lachin, J.M. (2000). *Biostatistical Methods - The Assessment of Relative Risks.* Wiley Series in Probability and Statistics.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* 2nd edition. London: Chapman & Hall.

Calculating Estimates of an Effect in Stratified Nonparametric Analysis

Michael O'Kelly[1]

[1] Quintiles Ireland Ltd., East Point Business Park, Fairview, Dublin 3.

Abstract: This paper examines an overall estimate of effect based on the stratified Wilcoxon test.

Keywords: Nonparametric confidence interval; Hodges-Lehmann estimate; Wilcoxon test; Van Elteren test.

1 Introduction

Parametric models offer stratified estimates of overall treatment effect on the scale of the response. But there is no widely used method of estimating treatment effect associated with stratified nonparametric tests such as the stratified Wilcoxon (or van Elteren) test. There are many areas of research where such an estimate is warranted because a stratum-by-treatment effect is not expected. Clinical trials are a common source of such data, the stratum being investigator, hospital or country.

2 Data Analysed

Four sets of simulated data were used. Each set had 10 strata, and each stratum had 16 observations, 8 from each of two treatment groups (A and B), making 160 observations in all. The 10 strata each had a baseline, which was randomly assigned by sampling from a Normal distribution. The individual observations for each stratum were then calculated by adding to the stratum baseline a second random sample from a standard Normal distribution. Finally, if the treatment group was B, 0.5 was added to the observation. Thus the true treatment difference would be expected to be estimated as 0.5. The sets of data differed as follows:

- Set 1 had no contamination
- Set 2 had 20 percent of observations contaminated with a uniform (-5, -5.01) distribution
- Set 3 had 10 % of observations contaminated with a high-variance Normal distribution (mean zero, standard deviation (SD)=10)

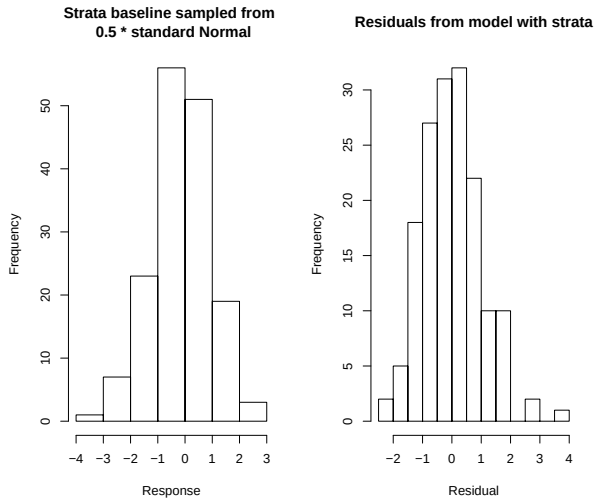


FIGURE 1. *Distribution of a single sample (160 observations) from data set 1*

See Figure 1 and Figure 2 for histograms illustrating these data. The strata for these first three sets had baselines sampled from a Normal distribution with mean zero and SD=0.5. Set 4 was as set 1, except that the strata baselines were more dispersed, being sampled from a standard Normal distribution. One thousand samples of each set were created.
In addition, analyses of highly-skewed mercury concentrations in Periphyton (Walpole and Myers, 1985, cited in Helsel and Hirsch, 2003) are briefly discussed.

3 Approaches Used

This paper proposes an extension of unstratified nonparametric methods of estimation (Lehmann, 1975)to stratified nonparametric tests such as the van Elteren test (van Elteren, 1960). Such extensions are not widely used or explicitly described in standard textbooks. An estimate based on the van Elteren test adapts the technique described by e.g. Sprent and Smeeton, (2001, pp. 1-43). This method was first suggested to the author by Kevin Kane of the statistical software company Phastar, and its rationale clarified in discussions with Dennis Boos and Gary Koch. In brief, the estimate of treatment difference is calculated by incrementing the response of one of the treatment groups until the expected value of the nonparametric statistic under the null hypothesis is found.The estimate of treatment difference is equal to the amount of the increment.In detail, the steps are as follows:

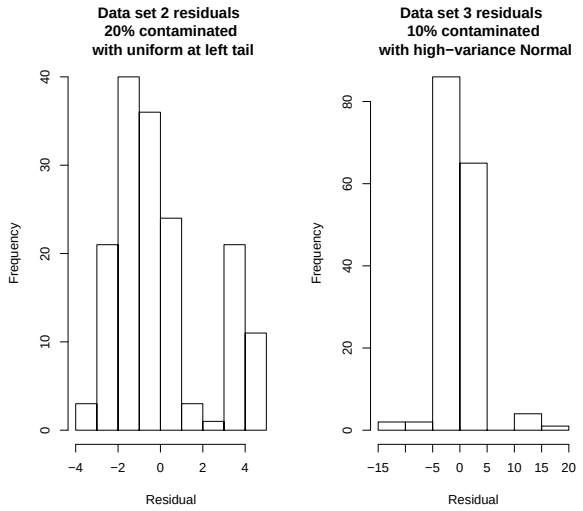


FIGURE 2. *Distribution of a single sample from data sets 2 and 3*

1. Using the Hodges-Lehmann estimate of median treatment difference, identify which treatment group has the lower median difference. Let us call this treatment group A.

2. Using as a start and end point the Hodges-Lehmann lower and upper CIs respectively, increase the value of all observations for group A by small increments, calculating the van Elteren test statistic after each increment.

3. At the point where the test statistic has reached its expected value under the null hypothesis of no treatment difference, the increment estimates the treatment difference under the van Elteren test.

The estimate of upper CI for the treatment difference is calculated similarly:

1. Note that the lower CI based on the van Elteren statistic will be greater than or equal to the lower CI of Hodges-Lehmann. Using as a start and end point the Hodges-Lehmann lower CI and the van Elteren estimate of treatment difference respectively, increase the value of all observations for group A by small increments, calculating the van Elteren test statistic after each increment.

2. At the point where the test statistic has reached the value which yields a two-sided significance of 5 percent,the increment estimates the lower CI for the treatment difference under the van Elteren test.

TABLE 1. *No contamination, strata baseline sampled from 0.5 * standard Normal*

Type of estimate	Cover	Power	CI length
One-way ANOVA	0.967	0.861	0.679
ANOVA stratified	0.940	0.903	0.595
Wilcoxon	0.966	0.828	0.706
Wilcoxon stratified	0.946	0.893	0.613

The upper CI is estimated in a similar manner to the lower CI. ANOVA-based estimates are used in this paper as comparators for the nonparametric estimates. The paper thus presents four estimates:

- As a baseline, the mean difference and its 95 percent CI based on the t-test (One-way ANOVA)
- Estimates from stratified ANOVA
- Nonparametric unstratified estimate based on the Wilcoxon test the Hodges-Lehmann estimate of median difference
- Nonparametric estimate based on the stratified Wilcoxon, or van Elteren, test

The coverage and power associated with the CI are presented below for estimates of treatment difference over 1000 instantiations of the four sets of data described above.

4 Main results

Table 1 acts as a baseline and summarises the results for the four methods of estimation over 1000 simulations for data with strata differing in their overall mean, but otherwise Normally distributed with a true treatment difference of 0.5. Results are as expected, with coverage close to the nominal coverage and lower power associated with the nonparametric CIs (Hodges and Lehmann, 1962).
However, when the disparity between the strata increases so that the stratum baseline is sampled from a standard Normal distribution, the unstratified Wilcoxon (Hodges-Lehmann) estimate loses power considerably (Table 2)
Table 3 features simulations with 20 percent of the data concentrated in the left tail of the distribution. This pattern is often found in biometric measures. While coverage is largely unaffected for all estimates, the power of the estimates based on ANOVA is of course dramatically reduced; and the stratified nonparametric CI has increased its advantage in power relative

TABLE 2. *No contamination, strata baseline sampled from standard Normal*

Type of estimate	Cover	Power	CI length
One-way ANOVA	0.985	0.685	0.838
ANOVA stratified	0.941	0.902	0.596
Wilcoxon	0.990	0.635	0.872
Wilcoxon stratified	0.944	0.871	0.644

TABLE 3. *Contaminated with observations at left tail, strata 0.5*standard Normal*

Type of estimate	Cover	Power	CI length
One-way ANOVA	0.956	0.298	1.375
ANOVA stratified	0.944	0.337	1.307
Wilcoxon	0.962	0.509	0.992
Wilcoxon stratified	0.940	0.658	0.797

TABLE 4. *With high-variance Normal contamination, strata 0.5*standard Normal*

Type of estimate	Cover	Power	CI length
One-way ANOVA	0.944	0.194	2.032
ANOVA stratified	0.925	0.210	1.951
Wilcoxon	0.957	0.655	0.851
Wilcoxon stratified	0.925	0.774	0.737

to its unstratified nonparametric equivalent.Finally, Table 4 follows McKean and Vidmar (1994) in examining contamination with highly-variable Normal data. As noted in the paper by McKean and Vidmar, the power of the estimates based on ANOVA is particularly badly affected by the high-variance contamination. The stratified nonparametric CI retains its superiority in power over its unstratified nonparametric equivalent and of course over the ANOVA-based estimates.

Analysis of the mercury data in Helsel and Hirsch (2003) shows that where the stratum effect is not significant, the estimate based on the stratified Wilcoxon test has no significant advantage over the traditional unstratified equivalent, the Hodges-Lehmann estimate: the CIs are almost identical. It should be noted that this will often be the case, and that where differences between the strata are modest, the standard Hodges-Lehmann estimate and CI is often an adequate estimate of treatment effect.

5 Conclusions

In estimating a treatment effect and CI in a stratified nonparametric setting, the unstratified Hodges-Lehmann estimate and CI may be adequate with regard to coverage and power, if the strata have overall means only moderately different from one another. As a rough guide, the means should be consistent with a Normal distribution with a SD about half that of the residuals from a full model (i.e. a model which includes stratum effects). With greater dispersion between the strata, a stratified estimate such as that presented here is preferable. When data is Normally distributed, the ANOVA estimate is of course the best choice, but its well-known loss of power in the face of contamination by more highly dispersed data or skewed data is confirmed again by this paper.
Further research is desirable to compare the nonparametric estimates proposed in this paper with those of robust regression (e.g. McKean and Vidmar, 1994).

Acknowledgments: This research was sponsored by Quintiles Ireland Ltd.. My thanks also to David Williams of University College Dublin for help with the formatting of this paper.

References

Helsel, D. R. and Hirsch, R.M. (2003). *Statistical Methods in Water Resources*. WWW: http://water.usgs.gov/pubs/twri/ (previously published edition dated 1992, Elsevier Science, now out of print)

Hodges, J.L. and Lehmann, E.L. (1962). Rank methods for combination of independent experiments in analysis of variance. *Annals of Mathematical Statistics*, **33**, 482–497.

Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Oakland: Holden-Day Inc.

McKean, J.W. and Vidmar, T.J. (1994). A comparison of two rank-based methods for the analysis of linear models. *The American Statistician*, **48**, 220–229.

Sprent, P. and Smeeton, N.C. (2001). *Applied Nonparametric Methods*, 3rd ed., London: Chapman & Hall.

van Elteren, P.H. (1960). On the combination of independent two-sample tests of wilcoxon. *Bulletin of the International Statistical Institute*, **37**, 351–361.

Walpole, R.E. and Myers, R.H. (1985). *Probability and Statistics for Engineers and Scientists*, 3rd ed. New York: MacMillan.

Harmonic Markov Switching Autoregressive Models for Bayesian Analysis of Air Pollution

Roberta Paroli[1] and Luigi Spezia[2]

[1] Istituto di Statistica, Università Cattolica S.C.- Milano
[2] Department of Statistics, Athens University of Economics and Business

Abstract: Markov switching autoregressive models (MSARMs) are efficient tools to analyse nonlinear and nongaussian time series. A special MSARM with a harmonic component is here proposed in the bayesian framework to analyse periodic time series. We perform a complete Gibbs sampling algorithm for model choice, for constraint identification and for the estimation of the unknown parameters and the latent data. We illustrate our methodology with two examples about the dynamics of air pollutants.

Keywords: Nonlinear and nongaussian time series; Latent variables; Carbon monoxide; Sulphur dioxide.

1 Introduction

Air quality control includes the study of data sets recorded by air pollution testing stations. We are interested in the analysis of the dynamics of the hourly mean concentrations of carbon monoxide (CO) and of the daily mean concentrations of sulphur dioxide (SO2). The main characteristics of the series that must be modelled are: *i*) different unobserved levels of pollutant mean concentrations, depending on the weather conditions (higher level of pollution in the colder days and lower in the warmer ones), *ii*) serially correlated data, *iii*) daily (CO) or yearly (SO2) periodicities, *iv*) missing observations. By these characteristics, *Markov switching autoregressive models* (MSARMs) (Hamilton (1994), ch. 22) can be efficient tools to analyse these environmental time series. A special MSARM with a harmonic component is here proposed in the bayesian framework, giving rise to Harmonic MSARMs (HMSARMs).

2 Harmonic Markov Switching Autoregressive Models

MSARMs of order $(m; p)$, henceforth MSAR(m;p), are discrete-time stochastic processes $\{Y_t; X_t\}$, such that $\{X_t\}$ is an unobservable discrete-time Markov chain with a finite number of states, m, while $\{Y_t\}$, given $\{X_t\}$, is

an observed autoregressive process of order p with the conditional distribution of Y_t depending on $\{X_t\}$ only through the contemporary X_t. Let $\{X_t\}$ be a discrete, first-order, homogeneous, ergodic Markov chain on a finite state-space S_X with cardinality m $(S_X = \{1, \ldots, m\})$. $\Gamma = [\gamma_{i,j}]$ is the $(m \times m)$ transition matrix, where $\gamma_{i,j} = P(X_t = j \mid X_{t-1} = i)$, for any $i, j \in S_X$, $x^T = (x_1, \ldots, x_T)'$ is the sequence of the states of the Markov chain and, for any $t = 1, \ldots, T$, x_t assumes values in S_X. Hence, given the order-p dependence and the contemporary dependence conditions, the equation describing HMSARMs is

$$Y_{t(x_t)} = \mu_{x_t} + \varphi_{1(x_t)} Y_{t-1(x_{t-1})} + \ldots + \varphi_{p(x_t)} Y_{t-p(x_{t-p})} + \eta_t + E_{t(x_t)}, \quad (1)$$

where $Y_{t(i)}$ denotes the generic variable Y_t when $X_t = i$, for any $1 \leq t \leq T$ and for any $i \in S_X$; the autoregressive coefficients $\varphi_{\tau(i)}$, for any $\tau = 1, \ldots, p$ and for any $i \in S_X$, depend on the current state i of the Markov chain; η_t is a harmonic component of periodicity $2s$,

$$\eta_t = \sum_{j=1}^{s^*} \left(\eta_{1,j} \cos\left(\pi j t / s\right) + \eta_{2,j} \sin\left(\pi j t / s\right) \right), \quad (2)$$

where s^* is the number of significant harmonics $(s^* \leq s)$; $E_{t(i)}$ denotes the gaussian random variable E_t when $X_t = i$, with zero mean and precision λ_i $\left(E_{t(i)} \sim \mathcal{N}\left(0; \lambda_i\right)\right)$, for any $i \in S_X$, with the discrete process $\{E_t\}$, given $\{X_t\}$, satisfying the conditional independence and the contemporary dependence conditions. Notice that the harmonic component does not depend on the hidden Markov chain for identifiability reasons: if it depended, to have an identified model, we would assume the same hidden state all along the period $2s$. The labels of the states and the sub-models, given a state, are interchangeable; the model (1) is unidentifiable in data fitting and therefore we need the following identifiability constraint: $\lambda_i < \lambda_j$, for any $i, j \in S_X$ such that $i < j$. In the following section we shall see how and why we choose this special constraint. At this point it is important only to notice the constraint is chosen *ex post* after simulations in such a way to respect the geometry and the shape of the unconstrained posterior distribution.

The parameters of the model, the latent data and the missing observations are estimated by simulation, performing Gibbs sampling, placing conjugate priors; the sequence of hidden states is estimated through the *forward filtering-backward sampling* algorithm by Carter and Kohn (1994) and Frühwirth-Schnatter (1994). Before performing parameter estimation we need to choose the best model and to select its identifiability constraints. MSAR model choice is performed through Bayes factor, computing the marginal likelihoods by the Chib-Neal method (Chib (1995), Neal (1999)). The selection of the identifiability constraints is done exploiting the mix-

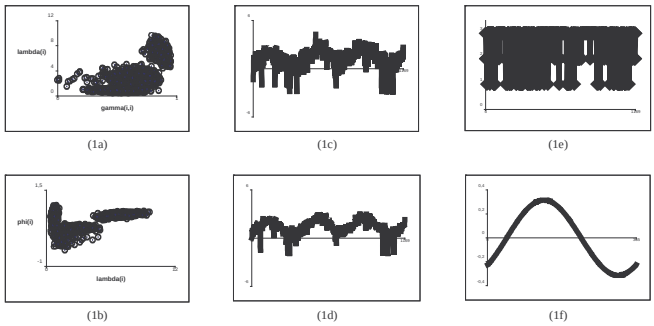


FIGURE 1.

ing properties of the *random permutation sampling algorithm* (Früwirth-Schnatter (2001)). In any iteration of the Gibbs sampler the entries of the vector λ must be in increasing order to satisfy the identifiability constraint. If λ is not ordered, instead of rejecting the vector and going on sampling till we have an ordered one, we introduce the *constrained permutation sampling algorithm* (Früwirth-Schnatter (2001)): we apply a permutation to order the precisions and consequently we apply the same permutation to the generated sequence of states and to the switching-parameters previously generated.

3 Applications to Air Pollution

Two applications of HMSARMs to real data will be studied in the following and two time series, the first about the daily mean concentrations of sulphur dioxide (SO2) and the second about the hourly mean concentrations of carbon monoxide (CO), will be analysed in detail. In each application we shall compare many competing models which differ for the cardinality of the state-space of the hidden Markov chain and for the order of the autoregressive process. We shall go on three consecutive steps: *i*) model selection, *ii*) constraint identification, *iii*) parameter estimation.

3.1 Application to Daily Mean Concentrations of SO2

The first periodic time series we consider is about the daily mean concentrations of SO2, in micrograms per cubic meter, recorded by the air pollution testing station placed in Via Goisis, Bergamo (Italy) from the 13th of September, 1996, to the 25th of November, 1999 (1169 observations). In the

series of SO2 a yearly periodicity is evident ($2s = 365$) and the number s^* of the harmonics is one.
Model selection, performed by means of *Bayes factors* in which the *marginal likelihoods*, i.e. the normalizing constants of the posterior densities, are computed according to Chib (1995), corrected by the relabeling of the hidden states (Neal (1999)), that the HMSAR(3;1) is the best among all the competing models.
Now we have to carefully identify the constraint which must respect the geometry and the shape of the unconstrained posterior distribution. Identifiability constraint is chosen by eye, looking at the graphs of the output of the unconstrained Gibbs sampling performed associated with *random permutation sampling* (Frühwirth-Schnatter (2001)): we plot couples of outputs of the estimates of the parameters obtained via unconstrained Gibbs sampling with random permutations of the hidden states; after that we check if there are groups corresponding to the different states and if these groups suggest special ordering in the labeling.
Random permutation sampling is an easy adjustment we introduce in the Gibbs sampler: at any iteration all the steps of Gibbs sampling run unconstrained; then we randomly generate one of $m!$ ways of labelling the states and consequently update the sequence of the hidden states and any switching-parameter according to the selected permutation of the states. *Random permutation sampling* allows us to explore the whole support of the posterior distribution, improving the mixing property of the sampler because the chain is free to move through the different subspaces, and encourages the moves from the current subspace to one of the other $(m! - 1)$. Graphically analysing the outputs of the unconstrained HMSAR(3;1) model, we choose the constraint on the precisions: $\lambda_1 < \lambda_2 < \lambda_3$ (Figures 1a and 1b). Now we can run *constrained permutation Gibbs sampling* for the HMSAR(3;1) model to estimate its parameters.
We obtain that SO2 yearly dynamics, described by the η_t's, respects that of the climatic conditions: higher levels of SO2 in the colder periods of the year and lower levels in the warmer ones (Figure 1f). Moreover we can see the dynamics of the fitted values (Figure 1d) respects the dynamics of the actual data, the natural logarithms of the observations (Figure 1c). Finally we are interested in the dynamics of the hidden states, representing the three different levels of pollution occured during the analysed period, which we can observe in Figure 1e, where we have the sequence of the posterior modes of any generated state x_t, for any $t = 1, \ldots, T$.

3.2 Application to Hourly Mean Concentrations of CO

The second periodic time series we consider is about the hourly mean concentrations of CO, in milligrams per cubic meter, recorded by the air pollution testing station placed in Via San Giorgio, Bergamo (Italy) from the 20th of October, 1998, 1 a.m., to the 8th of December, 1998, 12 p.m. (1200

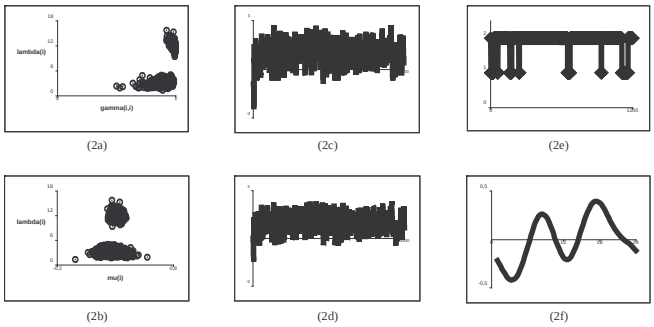


FIGURE 2.

observations). In the series of CO a daily periodicity is evident ($2s = 24$) and the number s^* of the harmonics is two. In the model choice phase we notice that the HMSAR(2;1) is the best among all the competing models. Then graphically analysing the outputs of theunconstrained HMSAR(2;1), we choose the constraint on the precisions again: $\lambda_1 < \lambda_2$ (Figures 2a and 2b). Finally, by the parameter estimation side, the CO daily dynamics, η_t's, respects the rush hours, in fact we have the peaks at eight a.m. and five p.m. (Figure 2f). The fitting performance of the model is evaluated through the plots of actual, the natural logarithms of the observations, and fitted values (Figures 2c and 2d): we have the fitted series well describes the observed phenomenon. By the Markov chain side, the estimated sequence of hidden states is plotted in Figure 2e.

4 Conclusions

The previously described empirical studies about air pollution show that Markov switching autoregressive models with a harmonic component well analyse periodic time series whose dynamics nonlinearly depend on latent variables. Model choice and inference have been performed through Gibbs sampling, considering the label switching problem, which has been efficiently tackled by permutation sampling.
The models we considered can be extended in many ways (i.e. time-varying transition matrices, multivariate pollutants and multisites recording analysis) to apply them more extensively to air quality control; these extensions are the subject of future researches.

References

Carter, C.K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, **81**, 541–553.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, **90**, 1313–1321.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, **15**, 183–202.

Frühwirth-Schnatter, S. (2001). Markov Chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, **96**, 194–209.

Hamilton, J.D. (1994). *Time Series Analysis*. Princeton University Press, Princeton.

Neal, R.M. (1999). Erroneous results in "marginal likelihood from the Gibbs output". *Unpublished manuscript*. http://www.cs.utoronto.ca/∼radford/.

Local Influence and Leverage in Elliptical Nonlinear Regression Models

Gilberto A. Paula[1], Francisco José de A. Cysneiros[2], and Manuel Galea[3]

[1] Instituto de Matemática e Estatística, USP - Caixa Postal 66281 (Ag. Cidade de São Paulo), 05311-970 São Paulo - SP - Brazil, Email: giapaula@ime.usp.br
[2] Departamento de Estatística, Universidade Federal de Pernambuco - Caixa Postal 50749-540, Recife - PE - Brazil, Email: cysneiros@de.ufpe.br
[3] Departamento de Estadística, Universidad de Valparaíso, - Casilla 5030, Valparaíso - Chile, Email: Manuel.Galea@uv.cl

Abstract: This work deals with the calculation of local influence curvatures and generalized leverage in univariate elliptical nonlinear regression 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 derive the total local influence of the ith observation C_i and we decompose the generalized leverage matrix into two terms, one that may be interpreted as a contribution of the position parameter estimates on the leverage and the other as a kind of correction due to the estimation of the dispersion parameter. This correction vanishes for the normal case. We also establish a connection between generalized leverage and local influence. An illustrative example is given.

Keywords: Elliptical distributions; Leverage; Likelihood displacement; Local influence; Residuals; Robust models.

1 Elliptical Nonlinear Regression Models

Let Y_i, $i = 1, \ldots, n$, be independent random variables with density function of the form

$$f_{y_i}(y_i) = \frac{1}{\sqrt{\phi}} g\{(y_i - \mu_i)^2/\phi\},\ y_i \in I\!R, \tag{1}$$

where $\phi > 0$ is the scale parameter, $g : I\!R \to [0, \infty]$ is such that $\int_0^\infty g(u^2)du < \infty$. We shall denote $Y_i \sim El(\mu_i, \phi)$. The function $g(.)$ is called density generator (see, for example, Fang, Kotz and Ng, 1990). The univariate elliptical nonlinear regression model is defined by

$$Y_i = \mu_i(\boldsymbol{\beta}) + \boldsymbol{\epsilon_i},$$

where $\mu_i(\boldsymbol{\beta}) = \mu(\boldsymbol{\beta}; \mathbf{x}_i)$ is a nonlinear function of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and $\boldsymbol{\epsilon_i} \sim El(0, \phi)$. An iterative process to get the maximum likelihood estimates

$\hat{\boldsymbol{\beta}}$ and $\hat{\phi}$ may be developed by using, for example, the scoring Fisher method. This joint iterative process is given by

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (4d_g)^{-1}\{\mathbf{D}_\beta^{(m)T}\mathbf{D}_\beta^{(m)}\}^{-1}\mathbf{D}_\beta^{(m)T}\mathbf{D}(\mathbf{v}^{(m)})\{\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta}^{(m)})\} \tag{2}$$

and

$$\phi^{(m+1)} = \frac{1}{n}Q_V(\boldsymbol{\beta}^{(m+1)}) \ \ (m = 0, 1, 2, \ldots), \tag{3}$$

where $Q_V(\boldsymbol{\beta}) = \{\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})\}^T\mathbf{D}(\mathbf{v})\{\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})\}$, $\mathbf{D}_\beta = \partial\boldsymbol{\mu}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$, $d_g = E\{W_g^2(U^2)U^2\}$ with $U \sim El_n(0,1)$, $W_g(u) = g'(u)/g(u)$ with $g(u) = \partial g(u)/\partial u$ and $\mathbf{D}(\mathbf{v}) = \text{diag}\{v_1, \ldots, v_n\}$ with $v_i = -2W_g(u_i)$ and $u_i = (y_i - \mu_i)^2/\phi$, $i = 1, \ldots, n$. We should start the iterative process (2)-(3) with initial values $\boldsymbol{\beta}^{(0)}$ and $\phi^{(0)}$.

2 Local Influence

Let $\text{L}(\boldsymbol{\theta})$ denote the log-likelihood function from the postulated model where $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \phi)^T$ and let $\boldsymbol{\omega}$ be a $n \times 1$ vector of perturbations restricted to some open subset $\boldsymbol{\Omega} \subset I\!R^n$. The perturbations are made on the likelihood function, such that it takes the form $\text{L}(\boldsymbol{\theta}|\boldsymbol{\omega})$. Denoting the vector of no perturbation by $\boldsymbol{\omega}_0$, it is assumed that $\text{L}(\boldsymbol{\theta}|\boldsymbol{\omega}_0) = \text{L}(\boldsymbol{\theta})$. The idea of local influence (Cook, 1986) is concerning with characterizing the behaviour of the likelihood displacement $\text{LD}(\boldsymbol{\omega}) = 2\{\text{L}(\hat{\boldsymbol{\theta}}) - \text{L}(\hat{\boldsymbol{\theta}}_\omega)\}$ around ω_0. It may be showed that the normal curvature at the direction $\boldsymbol{\ell}$ takes the form $\text{C}_{\boldsymbol{\ell}}(\boldsymbol{\theta}) = 2|\boldsymbol{\ell}^T\boldsymbol{\Delta}^T(\ddot{\mathbf{L}})^{-1}\boldsymbol{\Delta}\boldsymbol{\ell}|$ where $-\ddot{\mathbf{L}}$ is the observed Fisher information matrix for the postulated model ($\boldsymbol{\omega} = \boldsymbol{\omega}_0$) and $\boldsymbol{\Delta}$ is the $(p+1) \times q$ matrix with elements $\Delta_{ij} = \partial^2\text{L}(\boldsymbol{\theta}|\boldsymbol{\omega})/\partial\theta_i\partial\omega_j$, evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and $\boldsymbol{\omega} = \boldsymbol{\omega}_0$, $i = 1, \ldots, p+1$ and $j = 1, \ldots, n$. As perturbation scheme we shall consider the heteroscedastic model

$$f_{y_i}(y_i|\omega_i) = \sqrt{\frac{\omega_i}{\phi}}g\{\omega_i(y_i - \mu_i)^2/\phi\}, \tag{4}$$

with ω_i denoting the weight corresponding to the ith case, $i = 1, \ldots, n$. When $\omega_i = 1$, the perturbed model (4) reduces to the postulated model (1). Indeed, we are perturbing the scale parameter by changing it to ϕ/ω_i for the ith observation. We obtain $\boldsymbol{\Delta}^T = [-\frac{2}{\hat{\phi}}\mathbf{D}(\hat{\mathbf{b}})\mathbf{D}_{\hat{\beta}}, -\frac{1}{\hat{\phi}^2}\mathbf{D}(\hat{\mathbf{b}})\hat{\mathbf{e}}]$, that is an $n \times (p+1)$ matrix, where $\mathbf{D}(\mathbf{b}) = \text{diag}\{b_1, \ldots, b_n\}$ with $b_i = \{W_g(u_i) + u_iW_g'(u_i)\}e_i$ and $e_i = y_i - \mu_i(\boldsymbol{\beta})$, $i = 1, \ldots, n$. Then, the total local influence of the ith observation (Lesaffre and Verbeke, 1998) yields $\text{C}_i = 2|\boldsymbol{\ell}_i^T\boldsymbol{\Delta}^T(\ddot{\mathbf{L}})^{-1}\boldsymbol{\Delta}\boldsymbol{\ell}_i|$, where $\boldsymbol{\ell}_i$ is an $n \times 1$ vector of zeros with one at the ith position. We can express $-\ddot{\mathbf{L}}$ in a closed-form expression. Attention should be given to those observations with $\text{C}_i > 2\bar{C}$.

3 Generalized Leverage

Let $\hat{\mathbf{y}} = \boldsymbol{\mu}(\hat{\boldsymbol{\beta}})$ be the predicted response vector. The main idea behind the concept of leverage (see, for instance, Emerson, Hoaglin and Kempthorne, 1984; St. Laurent and Cook, 1992; Wei, Hu and Fung, 1998) is that of evaluating the influence of y_i on its own predicted value. This influence may be well represented by the derivative $\partial\hat{y}_i/\partial y_i$ that equals h_{ii} in the normal linear case, where h_{ii} is the ith principal diagonal element of the projection matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ and $\mathbf{X}$ is the model matrix. Extensions to more general regression models have been given, for instance, by St. Laurent and Cook (1992) and Wei et al.(1998). Using equation (2.6) of Wei et al. (1998) the $(n \times n)$ matrix $(\partial\hat{\mathbf{y}}/\partial\mathbf{y})$ of generalized leverage in univariate elliptical nonlinear regression models may be expressed as

$$\mathbf{GL}(\hat{\boldsymbol{\theta}}) = \mathbf{GL}_{\beta}(\hat{\boldsymbol{\theta}}) + \mathbf{GL}_{\phi}(\hat{\boldsymbol{\theta}}),$$

where $\mathbf{GL}_{\beta}(\hat{\boldsymbol{\theta}})$ may be interpreted as a contribution of $\hat{\boldsymbol{\beta}}$ on the leverage while $\mathbf{GL}_{\phi}(\hat{\boldsymbol{\theta}})$ is a kind of correction due $\hat{\phi}$. In particular, for the normal case, the generalized leverage matrix $\mathbf{GL}(\hat{\boldsymbol{\theta}})$ reduces to the Jacobian leverage matrix

$$\hat{\mathbf{J}} = \mathbf{D}_{\hat{\beta}}\left\{\mathbf{D}_{\hat{\beta}}^T\mathbf{D}_{\hat{\beta}} - [\hat{\mathbf{e}}^T][\mathbf{D}_{\hat{\beta}\hat{\beta}}]\right\}^{-1}\mathbf{D}_{\hat{\beta}}^T. \tag{5}$$

St. Laurent and Cook (1992) compare (5) with the tangent plane leverage matrix $\hat{\mathbf{H}} = \mathbf{D}_{\hat{\beta}}(\mathbf{D}_{\hat{\beta}}^T\mathbf{D}_{\hat{\beta}})^{-1}\mathbf{D}_{\hat{\beta}}^T$, that is the orthogonal projection matrix onto the subspace spanned by the columns of the matrix $\mathbf{D}_{\hat{\beta}}$. If we use the perturbation scheme $y_{i\omega_i} = y_i + \omega_i$ and we assume ϕ fixed then $\mathrm{C}_i = \frac{|\hat{a}_i|}{\hat{\phi}}\mathrm{GL}_{ii}$, where $a_i = -2\{W_g(u_i) + 2u_iW_g'(u_i)\}$.

4 Application

In order to illustrate an application we shall consider the data set described in Ratkowsky (1983, Table 6.1) on the weight of the dried eye lens, Y (mg) of the European rabbit *Oryctolagus cuniculus* versus the age of the animal, X (days), a sample of 71 observations. This animal is largely distributed in wild populations in Australia. A three-parameter model that presented both intrinsic and parameter-effects curvatures non-significant under normal error with constant variance for log$Y$, shall also be considered here under other elliptical errors. An interesting aspect of this data set that supports the use of error distributions with heavier tails than the ones of the normal distribution is the suspicion of two outliers under least-squares estimation. Then, to reanalyze the data, we propose the following model:

$$Y_i = \exp\left(\alpha - \frac{\beta}{x_i + \gamma}\right)e^{\epsilon_i},$$

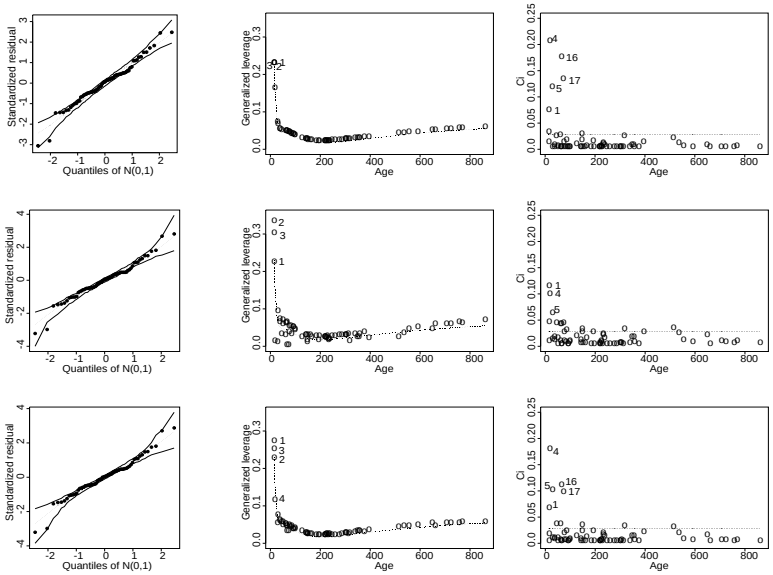


FIGURE 1. *Envelopes and index plots of* GL_{ii} *and* C_i *for the normal (top), Student-t with 10 df (middle) and logistic-II (bottom), fitted on the rabbit data.*

where $\epsilon_i \sim El(0, \phi)$ are mutually independent errors. A Student-t model with 10 degrees of freedom and a logistic-II model were also fitted to the data. The maximum likelihood estimates do not differ much among the three fitted models, but the approximate standard errors of the Student-t and logistic-II models are smaller than the ones of the normal model. Figure 1 presents some diagnostic graphics. Even though observations 16 and 17 appear as possible outliers in all the fitted models the generated envelopes do not present any unusual features. Observations 1, 2 and 3 appear as high leverage points in the three models. The Student-t model stands out less observations in the index plot of $\mathrm{C_i}$ than the logistic-II and normal models. We can notice from these graphics that younger animals tend to be more influential on the parameter estimates and on their own fitted values. The dotted lines in the graphics of GL_{ii} represent the index plot of $\hat{h}_{ii}$ (tangent plane leverage) which are negligible, as expected, for the normal case, but differ for the outstanding observations in the Student-t and logistic-II models. Elimination of the observations 16 and 17 produces larger changes in the estimates of the normal model than in the estimates of

the Student-t and logistic-II models. However, elimination of the influential and high leverage points does not change much the parameter estimates but produces considerable changes in the approximate standard errors.

Acknowledgments: The first author received financial support from CNPq and FAPESP, Brazil, the second author was supported by CAPES, Brazil and the third author by FONDECYT, Chile.

References

Cook, R.D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society, Series B*, **48**, 133–169.

Emerson, J.D., Hoaglin, D.C., and Kempthorne, P. J. (1984). Leverage in least squares additive-plus-multiplicative fits for two-way tables. *Journal of the American Statistical Association*, **79**, 329–335.

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

Lesaffre, E. and Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics*, **38**, 963–974.

Ratkowsky, D.A. (1983). *Nonlinear Regression Modeling*. New York: Marcel Dekker.

St. Laurent, R.T. and Cook, R.D. (1992). Leverage and superleverage in nonlinear regression. *Journal of the American Statistical Association*, **87**, 985–990.

Wei, B.C., Hu, Y.Q., and Fung, W.K. (1998). Generalized leverage and its applications. *Scandinavian Journal of Statistics*, **25**, 25–37.

Self-consistent Partitioning of Functional Data for Profiling Placebo Responders

Eva Petkova[1], Thaddeus Tarpey[2], and Todd Ogden[1]

[1] Columbia University, Department of Biostatistics, 1051 Riverside Dr. Unit 48, New York, NY 10025, USA
[2] Wright State University, Department of Mathematics and Statistics, Dayton, Ohio 45435, USA

Abstract: Identification of placebo responders among subjects treated with active drug has significant clinical and research implications. In clinical practice when a patient treated with medication improves, this improvement may be attributed to the chemical component of the drug itself ("true drug effect"), a "placebo effect", or some combination of these. Determining the proper subsequent treatment and maintenance of the patient may be aided greatly by understanding the type of a patient's response. This work presents a framework for studying placebo response in diverse areas of medicine. In order to identify placebo responders among drug treated patients, a profile of the clinical status over time (outcome profile) is estimated for each subject. Self-consistent partitioning techniques are used to group subjects based on the amount of curvature in the profile as well as the overall trend in the profile. The resulting partitions determine representative profiles for subjects in the drug group which can subsequently be used to classify patients. The method is applied to data from a clinical trial for treatment of depression involving placebo and the active drug phenelzine.

Keywords: Clustering; Principal points; Self-consistent points; Specific drug response.

1 Introduction and Background

Identifying placebo responders among drug-treated patients is an important problem for practicing clinicians and is at the heart of numerous long standing issues in drug research, Kahn and Brown, 2001. In psychiatry, ill people who are treated and improve are called responders. Responders treated with active drug may have improved due to a *true drug effect* or they may have responded to non-specific aspects of the treatment, called *placebo effect*. For the purposes of this paper *placebo effect* is defined as the totality of effects that cannot be attributed to the active chemical component of the drug, such as the effect of taking a pill and interacting with and receiving attention from clinicians and nurses. *True drug effect* is defined to represent the effect of the active chemical compound in the drug that is

not contained in the placebo pill. Clinical decisions will be affected if patients on drugs can be identified as achieving a placebo effect or achieving a true drug response. For example, in standard clinical practice, subjects identified as *placebo responders* may not need a continued drug treatment. In addition, such patients might require more frequent observation by the treating clinician since they are at higher risk for relapse than patients who experience a true drug response (Stewart et al, 1998).
In antidepressant studies, response rates among placebo-treated subjects are substantial and can range from 25% to 40%, Kahn and Brown, 2001. Clearly, among responders in the active treatment group there will be *placebo responders* as well as *true drug responders*. Quitkin et al (1997) describe certain patterns in the trajectory over time of the severity of depressive symptoms of patients treated with active drug and conclude that they are likely to correspond to *placebo effects* because they occur no more often on drug than on placebo, while others might represent *true drug responses*.
An established view in psychiatry gives rise to a classification of subjects treated with active drug into five non-overlapping categories presented below Quitkin *et al.*, 1997. Subjects treated with placebo can only fall only into one of the first three categories. (A.) *Non-responders:* subjects who do not improve throughout the trial. (B.) *Non-responders with initial placebo effect:* subjects who temporarily improve in the beginning of the treatment due to a non-specific effect, but deteriorate by the end of the study. (C.) *Placebo responders:* subjects who respond due to the non-specific effects of the treatment. (D) *True drug responders:* subjects who have a specific effect, i.e., subjects who respond to the active chemical component of the drug and not to any of the non-specific components of the treatment. (E.) *Mixture effect responders:* subjects whose final outcome is a combination of specific and non-specific effects; such are subjects who have an initial improvement due to non-specific effects and then experience a true drug effect.
The primary goal of this paper is to determine a way of identifying the placebo responders in the drug-treated group of patients.

2 Data Description

Data from a clinical trial for the treatment of depression are analyzed. Subjects were randomized to either the antidepressant phenelzine or to a placebo. The outcome measure for each subject was an integer score between 0 and 23 on the Hamilton Depression (Ham-D) scale, assessed at baseline (week 0) and then once a week for six weeks. Higher scores on the Ham-D indicate greater severity of depression.

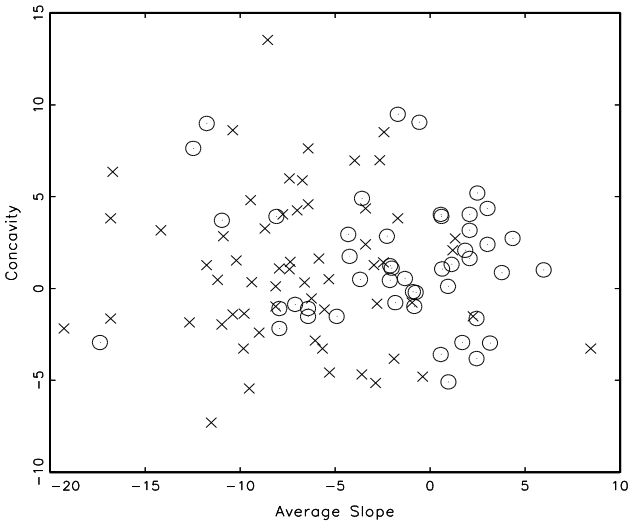


FIGURE 1. *A scatterplot of the estimated average-slope coefficients and concavity coefficients for the phenelzine (×) and placebo (⊙) groups.*

3 Functional Profile Modeling

The depression severity over time for each subject is modeled using a functional data analysis approach Ramsay and Silverman, 1997. The Ham-D response was modeled as a quadratic function of time plus a random error. Orthonormal basis functions are used to represent the functional data. Thus, the model for the ith individual is

$$y_i(t) = \beta_{0i} f_1(t) + \beta_{1i} f_1(t) + \beta_{2i} f_2(t) + \epsilon_i(t)$$

where the functions f_0, f_1, f_2 are constant, linear and quadratic respectively defined so that they are orthonormal over the range $0 \leq t \leq 6$ weeks.
Figure 1 shows a scatterplot of the β_1 and β_2 parameters for the active treatment group (×) and placebo group (⊙).

4 Clustering Functional Data

The average-slope and concavity coefficients are used to classify individuals in order to determine representative profiles for categories A - E in the Introduction. The classification is based on identification of clusters in the bivariate distribution of (β_1, β_2). Several clustering approaches are

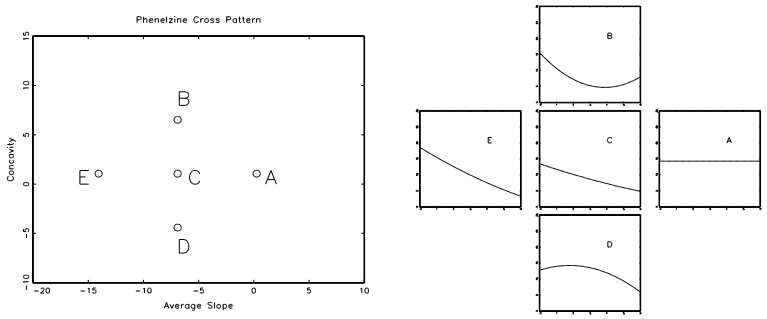


FIGURE 2. *The left frame shows the $k = 5$ cluster means from the constrained cross pattern with center point labeled by A, B, C, D, and E corresponding to the five outcome profile categories. These five cluster means are plotted as parabolas for the phenelzine group. The five curves shown here are the five representative curves for the phenelzine group.*

considered, among them the distribution-free k-means algorithm and ML assuming normality for the joint distribution of β_1 and β_2. The k-means algorithm which is designed to find distinct subgroups performs inefficiently, furthermore it typically produces numerous solutions depending on initial starting values which makes the interpretation of any single solution as the set of representative profiles problematic, Tarpey, 1998. The ML estimators are generally impractical to compute.

The approach taken is to classify drug treated subjects using methodology based on *principal points*, Flury, 1990. and *self-consistent points*, Flury, 1993, which is similar in spirit to learning vector quantization and reference point logistic classification. The set of principal points is the optimal k-point representation of a theoretical distribution in terms of mean squared error (MSE). Formally, a set of k points $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k$ are principal points for a random vector $\boldsymbol{X}$ if

$$E(\min_{j=1,\ldots,k} \|\boldsymbol{X} - \boldsymbol{\xi}_j\|^2) \leq E(\min_{j=1,\ldots,k} \|\boldsymbol{X} - \boldsymbol{y}_j\|^2)$$

for every set of k points $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$. The optimal one-point representation of a distribution (in terms of mean squared error) is the mean which corresponds to $k = 1$ principal point. Thus, principal points are simply a generalization of the mean from one to several points which optimally represent the distribution.

Symmetric multivariate distributions often have many different sets of self-consistent points. Tarpey, Tarpey, 1998, showed that the principal points (as well as other sets of self-consistent points) form symmetric patterns for the multivariate normal and other symmetric multivariate distributions. With the covariance configuration present in the data, a nearly optimal (in terms of MSE) cluster point pattern for the bivariate normal distribution is

TABLE 1. *Estimated classification counts and percentages for the phenelzine and placebo groups*

Frame	Description of outcome	Number in phenelzine group	Number in placebo group
A	Non-Responders	10 (16.9%)	30 (63.8%)
B	Non-Responders with Initial Placebo Effect	13 (22.0%)	6 (12.8%)
C	Placebo Responders	16 (27.1%)	8 (17.0%)
D	True Drug Responders	10 (16.9%)	1 (2.1%)
E	Mixture Effect Responders	10 (16.9%)	2 (4.3%)

the *cross pattern* with a center point, Tarpey, 1998. An efficient alternative to maximum likelihood is to find semiparametric estimates of self-consistent points constrained to form symmetric pattern susing grid search.

5 Results

The cross pattern (with center point) appears to perform best in terms of PMSE compared to other symmetric patterns and the solutions from the k-means algorithm. Figure 2 shows the constrained cross pattern cluster means plotted as parabolas in function space. A summary of the representative profiles are provided in Table 1 along with the corresponding counts (and percentages in parentheses) of subjects that fall in each of the categories.

References

Flury, B. (1990). Principal points. *Biometrika*, **77**, 33–41.

Flury, B. (1993). Estimation of principal points. *Applied Statistics*, **42**, 139–151.

Khan, A. and Brown, W.A. (2001). The placebo enigma in antidepressant clinical trials. *Journal of Clinical Psychopharmacology*, **21**, 123–125.

Quitkin, F.M., Rabkin, J.D., Markowitz, J.M., Stewart, J.W., McGrath, P.J., and Harrison W. (1997). Use of pattern analysis to identify true drug response. *Archives of General Psychiatry*, **44**, 259–264.

Ramsay, J.O. and Silverman, B.W. (1997). *Functional Data Analysis.* New-York: Springer.

Stewart, J.W., Quitkin, F.M., McGrath, P.J., Amsterdam, J., Fava, M., Fawcett, J., Reimherr, F., Rosenbaum, J., Beasley, C., and Roback, P. (1998). Use of pattern analysis to predict differential relapse of remitted patients with major depression during 1 year of treatment with fluoxetine and placebo. *Archives of General Psychiatry*, **55**, 334–343.

Tarpey, T. (1998). Self-consistent patterns for symmetric multivariate distributions. *Journal of Classification*, **15**, 57–79.

Applications of Some Characterizations for Count Data Distributions

Pedro Puig[1] and Jordi Valero[2]

[1] Servei d'Estadística. Universitat Autònoma de Barcelona, Spain.
[2] Escola Superior d'Agricultura de Barcelona. Universitat Politècnica de Catalunya, Spain.

Abstract: In this paper we find all two-parameter count distributions (satisfying very general conditions) that are closed under addition so that their maximum likelihood estimator of the population mean is the sample mean. For count distributions only additively closed with respect the population mean, the behaviour of their proportion of zeroes characterizes the distribution. Several examples of application of these results are commented.

Keywords: Overdispersion; Zero-inflation; Closed-under-addition; Hermite distribution.

1 On the Generalized Hermite Distribution

The Poisson distribution is arguably the most widely used distribution in modeling count data. There are a variety of reasons for this, including principles (like the law of rare events) that suggest how and why it arises so frequently in applications as well as certain properties which facilitate its use. One such property is its closure under addition, whereby sums of independent Poissons are again Poisson-distributed. Another is the fact that the sample mean is the maximum likelihood estimator of the mean of a Poisson-distributed population.
To model departures from Poisson distribution that produce situations like overdispersion or zero inflation it is reasonable to consider discrete distributions with more than one parameter. The following result characterizes, under very general conditions, all two-parameter count distributions so that they are closed under addition and their maximum likelihood estimator of the population mean is the sample mean (Puig, 2003):

Theorem 1: Given a count variate X that can be parameterized by its mean μ and variance σ^2, with a pgf continuous in μ and σ^2, closed under convolutions so that the maximum likelihood estimator of μ is the sample mean. Then the distribution of X is the same as $n_1Y_1 + n_2Y_2$, where Y_i are two independent Poisson variates with means $(\mu n_2 - \sigma^2)/(n_1(n_2 - n_1))$ and $(\sigma^2 - \mu n_1)/(n_2(n_2 - n_1))$ respectively, where n_1 and n_2 are positive integers, $n_1 < n_2$.

From the practical point of view, not all distributions characterized by this theorem are useful. If $n_1 > 1$ then some positive integers can never occur. For instance, if $n_1 = 2$ and $n_2 = 5$, then the values 1 and 3 have a probability equal to 0. For this reason we only consider the situation where $n_1 = 1$ and $n_2 = n$ and it corresponds to a family of distributions known as Generalized Hermite distribution (Gupta and Jain, 1974). When $n = 2$ this is known as Hermite distribution, which was introduced by Kemp and Kemp (1965, 1966).
By using this family of distributions we have analyzed the daily death registers for men and women aged 95 years and over in the Comunidad Autnoma de Madrid during 1995 (source from the Centre d'Estudis Demogrfics). Some properties of these data sets justify to use a two parameter count distribution closed under addition (see Puig, 2003).

2 Zero Inflation and Overdispersion

Definition: Let X denote a non-negative integer random variable (count variable) such that its mean is μ and its proportion of zeroes is p_0. The zero inflation index of X is, $zi(X) = 1 + \log(p_0)/\mu$.

Notice that zi(X)=0 if X is Poisson distributed and $zi(X) > 0$ if X is 'zero inflated', that is, its proportion of zeroes is greater than the proportion of zeroes of a Poisson variate having the same mean.
Many of the random variables used to modelize count data are zero inflated and overdispersed. For instance, it happens for any mixture of Poisson distributions. Moreover, for two parameter count models that can be parameterized by their mean μ and dispersion index $d = V(X)/\mu$, often their zero inflation index is a function that only depends of d. Table 1 shows the log-pgf (log-probability generating function) and these functional relations for some of the most frequently employed count distributions.
Notice that all these distributions are closed under addition, if the parameter d is fixed for all the independent variates that are summed. Moreover, their maximum likelihood estimator of μ is also the sample mean.
The following result clarifies the importance of the relation between zero inflation and dispersion indexes:

Theorem 2: Let X be a count variate that can be parametrized by its mean μ and dispersion index d, with a pgf continuous in μ and twice differentiable with continuity in d. Suppose that X is closed under convolutions when d is fixed and the maximum likelihood estimator of μ is the sample mean. Then $zi(X) = f(d)$, for some appropriate real valued function $f(.)$, and this function characterizes X.

The proof is based on the paper of Sprott (1983).
This theorem can be applied in exploratory data analysis in order to choose and appropriate count data model, when the researcher analyzes several

TABLE 1. *Relation between zero-inflation and dispersion indexes for some discrete distributions parameterized by μ and d (Neg.Bin.=Negative Binomial, P.I.G.=Poisson-Inverse Gaussian).*

Name	**log-pgf**	ZI-index
Neg. Bin.	$-\frac{\mu}{d-1}\log(1-(d-1)(t-1))$	$1+\frac{\log(d)}{1-d}$
Neyman A	$\frac{\mu}{d-1}(e^{(d-1)(t-1)}-1)$	$\frac{e^{1-d}+d-2}{d-1}$
Polya-Aeppli	$\frac{2\mu(1-t)}{(d-1)(t-1)-2}$	$1-\frac{2}{d+1}$
Hermite	$\mu((d-1)(t^2-1)/2+(2-d)(t-1))$	$\frac{d-1}{2}$
P.I.G.	$\frac{\mu}{d-1}(1-\sqrt{1-2(d-1)(t-1)})$	$\frac{d-\sqrt{1+2(d-1)}}{d-1}$

TABLE 2. *Six frequency distributions of automobile insurance claims.*

	No. of claims							
Data set	0	1	2	3	4	5	6	7
1	103704	14075	1766	255	45	6	2	
2	370412	46545	3935	317	28	3		
3	7840	1317	239	42	14	4	4	1
4	3719	232	38	7	3	1		
5	96978	9240	704	43	9			
6	20592	2651	297	41	7	0	1	

samples coming from similar experiments and the observed proportion of zeroes is high. This is the situation for the examples that we have studied. For instance, the counts of microarthropods in several samples of forest soil, or the counts of chromosomic abnormalities in 5000 cells when they are bombed with different doses of radiation (source 'Departament de Biologia Animal, de Biologia Vegetal i d'Ecologia, UAB').

Theorem 2 suggests a simple way that can help us to decide which count distribution can be used to fit the overall data sets. The method is to draw scatter-plots with the estimated values of $zi(X)$ and d, and compare the observed profiles with the theoretical profiles of some count distributions like those showed in Table 1.

Theorem 2 also lets to construct new two parameter count distributions for a given relation $zi(X) = f(d)$. The following example illustrates these procedures.

2.1 An Example: Automobile Claim Data Sets.

Gossiaux and Lemaire (1981) analyzed six data sets giving the number of automobile insurance claims per policy over a fixed period of time. These

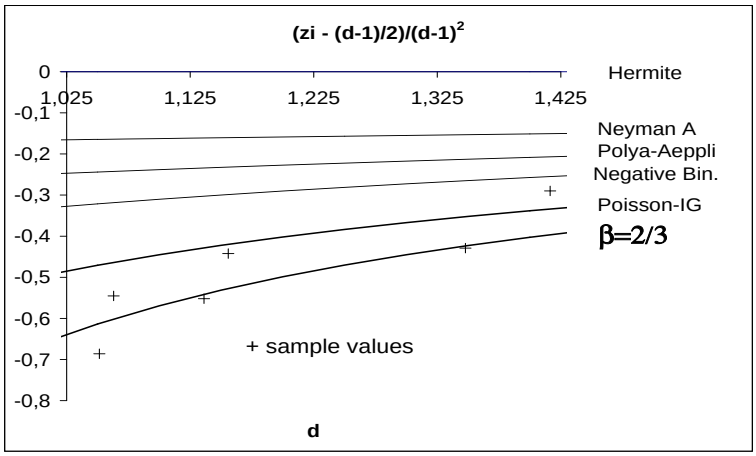


FIGURE 1. *Rescaled zi index versus d for some count distributions.*

data sets are shown in Table 2.
Poisson distribution is not adequate to fit these data sets because there are overdispersion. The Negative Binomial distribution improves the fits but G. E. Willmot (1987) remarks that the Poisson-Inverse Gaussian distribution works better.
Figure 1 shows the plots of the re-scaled zero-inflation index zi with respect to disperson index d for several of the count data distributions mentioned above. The sample values corresponding to the six data sets are also plotted. The plots show that the choice of the Poisson-Inverse Gaussian distribution is more adequate than the choice of the Hermite, Neyman A, Polya-Aeppli or Negative Binomial. However we can consider a new distribution with a log-pgf of the form $\mu(1-\beta)\frac{1-(1-(d-1)(t-1)/(1-\beta))^{\beta}}{\beta(d-1)}$, with $\beta = 2/3$.
Notice in Figure 1 how the performance of this new distribution is better than those obtained by using the Poisson-Inverse Gaussian.
The results of this graphical exploratory analysis coincide with the results of a goodness of fit chi-squared based analysis. Table 3 shows the chi-squared goodness of fit tests statistics with their corresponding p-values for the $\beta = 2/3$, Poisson-Inverse Gaussian and Negative Binomial distribution.
Observe that the $\beta = 2/3$ distribution provides p-values higher than Poisson-Inverse Gaussian, for 4 of the data sets and for the overall. It is clear that the Negative Binomial distribution provides poor fits.

TABLE 3. *Chi-squared goodness of fit test statistics for the six data sets. The values in brackets are the p-values.*

	Data set						
Distrib.	1	2	3	4	5	6	Overall
$\beta = 2/3$	3.675	0.455	2.816	2.178	5.635	0.226	14.983
	(.299)	(.929)	(.421)	(.536)	(.131)	(.973)	(.663)
P.I.G.	0.601	3.393	5.256	0.594	6.505	0.760	17.109
	(.896)	(.335)	(.154)	(.898)	(.089)	(.859)	(.516)
Neg. Bin.	14.104	9.916	18.524	1.408	9.272	3.887	57.111
	(.003)	(.019)	(.000)	(.704)	(.026)	(.274)	(.000)

References

Gossiaux, A. and Lemaire, J. (1981). Methodes d'ajustement de distributions de sinistres. *Bulletin of the Association of Swiss Actuaries*, **81**, 87–95.

Gupta, R.P. and Jain, G.C. (1974). A generalized Hermite distribution and its properties. *SIAM Journal on Applied Mathematics*, **27**, 359–363.

Kemp, C.D. and Kemp, A.W. (1965). Some properties of the 'Hermite' distribution. *Biometrika*, **52**, 381–394.

Kemp, A.W. and Kemp, C.D. (1966). An alternative derivation of the Hermite distribution. *Biometrika*, **53**, 627–628.

Puig, P. (2003). Characterizing additively closed discrete models by a property of their MLEs, with an application to generalized Hermite distributions. *Journal of the American Statistical Association*. (in press).

Sprott, D.A. (1983). Estimating the parameters of a convolution by maximum likelihood. *Journal of the American Statistical Association*, **78**(382), 457–460.

Willmot G.E. (1987). The poisson-inverse gaussian distribution as an alternative to the negative binomial. *Scandinavian Actuarial Journal*, 113–127.

How to Make a Causal Diagram for Sparse Vector Autoregression

Marco Reale

[1] Mathematics and Statistics Department, University of Canterbury, Private Bag 4800 Christchurch, New Zealand. Email: marco.reale@canterbury.ac.nz

Abstract: In this paper I present the procedure to construct a conditional independence graph for the variables included in a vector autoregression. I then derive the corresponding directed acyclic graph which has causality implication. I make use of the inflation transmission mechanism as an example.

Keywords: Causality; Graphical modeling; Vector autoregression.

1 Introduction

The relation among several autoregressions can be modeled with the vector autoregression

$$x_t = c + \Phi_1 x_{t-1} + \Phi_2 x_{t-2} + \ldots + \Phi_k x_{t-k} + e_t \tag{1}$$

of order k, VAR(k), where $x_t, x_{t-1}, \ldots, x_{t-k}$ are n-dimensional vectors with the corresponding coefficient vectors $\Phi_1, \Phi_2, \ldots, \Phi_k$, c is the constant and e_t is the error vector, which is assumed IID. If the covariance matrix, H, of e_t is not diagonal, the set of linear equations (1) corresponds to a system of seemingly unrelated regressions (Zellner, 1962) and in H are hidden the relations among the components of x_t. To highlight such relations we can represent the canonical VAR(k) in (1) in its structural form (SVAR) (Sims, 1986):

$$\Theta_0 x_t = d + \Theta_1 x_{t-1} + \Theta_2 x_{t-2} + \ldots + \Theta_k x_{t-k} + u_t \tag{2}$$

where $\Theta_i = \Theta_0 \Phi_i$ for $i = 0, \ldots, k$, $d = \Theta_0 c$ and $u_t = \Theta_0 e_t$ with covariance matrix $\Theta_0 H \Theta_0' = D$, which is diagonal.

If there are no zeros in the coefficient vectors, the SVAR is saturated but in many cases some lagged variables on the RHS in (2) do not play any role in explaining the current variables, x_t. In this case the value of the corresponding coefficient is zero and hence the SVAR is sparse.

In this paper we will identify sparse structures for vector autoregression by using graphical modeling and the final directed acyclic graph (DAG) will point at possible causal interpretations.

An examination of the covariance matrix of the variables involved, both current and lagged, can assist in identifying the sparse structure by the

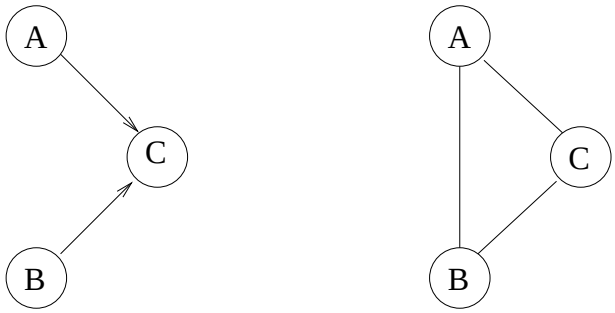


FIGURE 1. *Moralization of a directed acyclic graph.*

computation of the partial correlations using the *inverse variance lemma* (Whittaker, 1990, pp. 142–143). The significance of the partial correlations of model (2) can be tested using the appropriate sampling properties (Reale and Tunnicliffe Wilson, 2001 and 2002). In this way we obtain a conditional independence graph (CIG). The model (2) may be represented by a DAG in which the components of x_t, x_{t-1}, ..., x_{t-p} form the nodes, and causal dependence is indicated by arrows linking nodes.

Although the DAG and the CIG represent a different definition of the joint probability, there is a correspondence between these two graphs which is embodied by the moralization rule: because of this result we can obtain the CIG from the DAG by transforming the arrows into lines and linking unlinked parents. As a matter of example consider the graph in Figure 1: A and B are the *parents* of C. The moralization of the DAG on the left is obtained by transforming the existing arrows into edges and by adding and edge which links the parents. We define this kind of edges as *moral* edges. While the CIG represents the associations among the variables either in terms of conditional dependence or simply in terms of partial correlation, if the joint distribution is not Gaussian, the DAG has a natural interpretation in terms of causality. As it is not the aim of this paper to get involved in the philosophical debate around the definition of causality, we simply refer to a recent book by Pearl (2002).

The DAG is very attractive because of its causal interpretation but all we can observe in practice is the the CIG obtained by the sample partial correlation. So actually we need to perform the inverse operation of the moralization, we call it *demoralization*. Unfortunately while the transformation of a DAG into a CIG is unique, there are several DAG's which can give the same CIG. As an example consider the CIG on the right end side in figure 1: it could result from the moralization of all the DAG's in Figure 2. So we need to identify the moral links and remove them and to

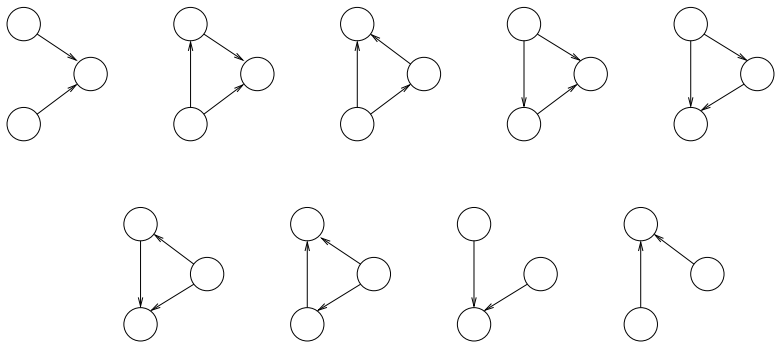


FIGURE 2. *Possible directed acyclic graph.*

do that we need to use all the knowledge we have about the relationships among the random variables in the system.
In the time series context the nature of the model is that all arrows end in nodes representing the contemporaneous variables on the left hand side of (2). Some arrows will start from the past, and some from other contemporaneous variables.
The coefficients are estimated by single equation ordinary least squares (OLS) regression. This is fully efficient under our working assumption, that the vector series is Gaussian. Our methods are also applicable, and the properties of the estimates given by the regression are reliable, under wider conditions, such as e_t being I.I.D., presented for example in Anderson (1971).

2 The Inflation Transmission

As a matter of example we apply our methodology to the inflation transmission between Italy (a), Germany (b), France (c) and the US (d) in the period January 1988 - December 2001.
We first identified a VAR of order 16 using the corrected Akaike information criterion and then used the *inverse variance lemma* to compute the partial correlation matrix for the variables $a(t), b(t), c(t), d(t), \ldots, a(t-16), b(t-16), c(t-16), d(t-16)$. We then test the significance of the partial correlation. As anticipated before we have a problem as we are testing several partial correlation simultaneously with a significant probability of making type I or type II errors. We use the strategy of testing using different levels of probability so we get a better feeling of the different significance.
In our specific case considered two levels of probability: 99% (bold lines)

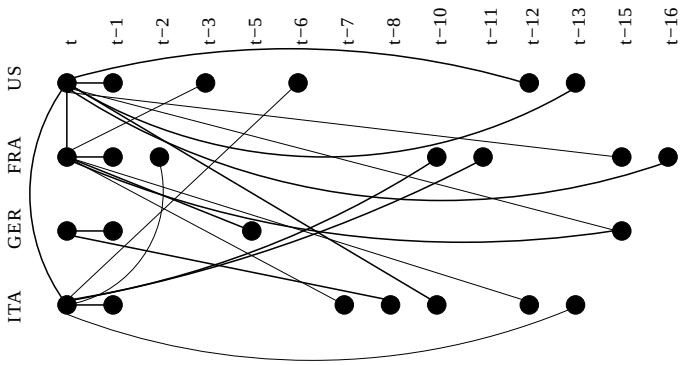


FIGURE 3. *Conditional independence graph.*

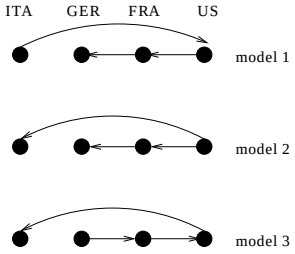


FIGURE 4. *Possible DAGs for current variables.*

and 95% (thin lines). The corresponding conditional independence graph is presented in Figure 3.
We then consider the possible DAG's for the current variables (see Figure 4) and apply subset regression in order to cancel moral links and derive the complete DAG's.
We eventually present the main alternative models and compare them using appropriate likelihood based methods similar to the AIC.

References

Anderson, T.W. (1971). *The Statistical Analysis of Time Series.* New York: Wiley.

Lauritzen, S.L. and Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their applications to expert

systems. *Journal of the Royal Statistical Society, Series B*, **50**, 157–224.

Pearl, J. (2000). *Causality*. Cambridge: Cambridge University Press.

Reale, M. and Tunnicliffe Wilson, G. (2002). The sampling properties of conditional graphs for structural vector autoregressions. *Biometrika*, **89**, 457–461.

Reale, M. and Tunnicliffe Wilson, G. (2001). Identification of vector AR models with recursive structural errors using conditional independence graphs. *Statistical Methods and Applications*, **10**, 49–65.

Sims, C.A. (1986). Are forecasting models usable for policy analysis? *Federal Reserve Bank of Minneapolis Quarterly Review*, **10**, 2–16.

Whittaker, J.C. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, **57**, 348–368

Power of Single Edge Exclusion in Graphical Log-Linear Models

M. Fátima Salgueiro[1], Peter W.F. Smith[2], and John W. McDonald[3]

[1] fatima.salgueiro@iscte.pt, Departamento de Métodos Quantitativos, ISCTE, Av. Forças Armadas 1649-026 Lisboa, Portugal
[2] pws@socsci.soton.ac.uk, Dept Social Statistics, University of Southampton, UK
[3] bigmac@socsci.soton.ac.uk, Dept Social Statistics, Univ of Southampton, UK

Abstract: The power of a backwards elimination model selection procedure is investigated for graphical log-linear models, with all variables binary. Asymptotic normal approximations are presented. Power calculations are illustrated using data on university admissions.

Keywords: Edge exclusion tests; Graphical log-linear models; Model selection; Odds ratios; Power.

1 Introduction

We investigate the power of a backwards elimination model selection procedure for graphical log-linear models (GLL) with two or three binary variables. We illustrate how to estimate power of single edge exclusion tests using asymptotic normal approximations. In Section 2 we review edge exclusion tests. In Section 3 we present normal approximations to the distributions of the likelihood ratio test statistic for single edge exclusion, under the alternative hypothesis that the saturated model holds. Results are used to approximate the power of the model selection procedure. Conclusions from a simulation study, to assess the quality of such approximations, are given. In Section 4 we illustrate power calculations using data on university admissions.

2 Edge Exclusion in Graphical Log-Linear Models

Graphical log-linear models are a subclass of hierarchical log-linear models, specified by setting a set of two-factor interaction terms λ_{ij} (and hence their higher-order relatives) to zero. The parameters of the GLL model are the remaining terms not set to zero. Such models can be interpreted solely in terms of conditional independence. Testing the null hypothesis that λ_{ij} and all higher-order interaction terms including it are zero is equivalent to

testing for conditional independence between the two corresponding factors X_i and X_j, given the remaining ones. For details see Whittaker (1990).

Let $X_1, \ldots, X_p$ be binary variables, here coded 0 and 1, and let $\pi_a(x_a)$ denote the marginal probability of $X_i = x_i : i \in a$. The total sample size equals $n_\emptyset$. Cell probabilities are assumed strictly positive, i.e., no structural zeros are allowed. Odds ratios are a commonly used measure of association in a contingency table. Let ψ_{ij} denote the marginal odds ratio between X_i and X_j (with i and j from 1 to p and $i \neq j$) and $\psi_{ij \cdot k}$ denote the conditional odds ratios, given a third binary variable X_k. A hat indicates a maximum likelihood estimator (m.l.e.).

Backwards elimination is a commonly used method for selecting a GLL model. The strategy is to start with the saturated model and test all the pairwise conditional independence statements, using test statistics for single edge exclusion. The likelihood ratio test (LRT) is the most commonly used test; alternatives include the Wald and the efficient score tests. Under the null hypothesis, each test statistic is asymptotically chi-squared distributed. In the two variables case signed square-root versions of the test statistics can also be used. Under the null, these are asymptotically standard normal distributed. The test statistics for single edge exclusion from a saturated GLL model are functions of many parameters (representing all higher order interaction terms), the number of parameters depending on the number of variables being considered. Hence, in general, the p variables case is complicated. For binary variables, Salgueiro (2002) presented closed form expressions for the test statistics, for $p = 2$ and 3, as a function of cell probabilities. Below only the non-signed version of the LRT statistics are considered.

In the two binary variables case $H_0 : \lambda_{12} = 0 \Leftrightarrow \psi_{12} = 1$ and the LRT statistic for the exclusion of edge (1,2) from the saturated GLL model is

$$L_{12} \;=\; 2\, n_\emptyset \sum_{x_1, x_2 \in \{0,1\}} \hat{\pi}_{12}(x_1, x_2) \log \left[\frac{\hat{\pi}_{12}(x_1, x_2)}{\{\hat{\pi}_1(x_1)\, \hat{\pi}_2(x_2)\}} \right]. \qquad (1)$$

With three binary variables, the LRT statistic for excluding edge (i, j) from the saturated GLL model ($i \neq j$, from 1 to 3), with $H_0 : \lambda_{ij} = \lambda_{ijk} = 0 \Leftrightarrow \psi_{ij \cdot k=0} = \psi_{ij \cdot k=1} = 1$, is

$$L_{ij} \;=\; 2 n_\emptyset \sum_{x_i, x_j, x_k \in \{0,1\}} \hat{\pi}_{ijk}(x_i, x_j, x_k)\; \log \left\{ \frac{\hat{\pi}_{ijk}(x_i, x_j, x_k)\, \hat{\pi}_k(x_k)}{\hat{\pi}_{ik}(x_i, x_k)\, \hat{\pi}_{jk}(x_j, x_k)} \right\}. \qquad (2)$$

3 Power of Single Edge Exclusion Tests

The test statistics for single edge exclusion from the saturated GLL model presented in Section 2 can be written as a function of the λ-terms of the

log-linear expansion. Let $\theta = vec(\lambda)$ be the vector of parameters of interest. Its m.l.e., based on $n_\emptyset$ observations, is $\hat{\theta} = vec(\hat{\lambda})$ and has an asymptotic normal distribution with mean θ and variance given by the inverse of the information matrix.

Salgueiro(2002) used the delta method to derive asymptotic normal approximations to the distributions of the test statistics for single edge exclusion from the saturated GLL model, under the alternative that the saturated model holds. As a result, L_{ij} is asymptotically normal distributed. For $p = 2$ and 3, respectively, the mean $AE[L_{ij}]$ is given by (1) and (2), with estimators replaced by parameters, and the variance $\text{var}(L_{ij})$ by

$$\text{var}(L_{12}) \;=\; 4n_\emptyset \sum_{x_1,x_2\in\{0,1\}} \pi_{12}(x_1,x_2)\log^2\left(\frac{\pi_{12}(x_1,x_2)}{\pi_1(x_1)\,\pi_2(x_2)}\right) - \frac{1}{n_\emptyset}\,(AE[L_{12}])^2,$$

$$\begin{aligned}\text{var}(L_{ij}) \quad &= \quad 4\,n_\emptyset \sum_{x_i,x_j,x_k\in\{0,1\}} \pi_{ijk}(x_i,x_j,x_k)\;\log^2\left(\frac{\pi_{ijk}(x_i,x_j,x_k)\,\pi_k(x_k)}{\pi_{ik}(x_i,x_k)\,\pi_{jk}(x_j,x_k)}\right)\\ &\quad -\frac{1}{n_\emptyset}(AE[L_{ij}])^2.\end{aligned}$$

For $p = 3$,

$\text{cov}(L_{ij}, L_{ik}) = -\frac{1}{n_\emptyset}\,(AE[L_{ij}])\,(AE[L_{ik}]) + 4\,n_\emptyset$

$$\sum\left[\pi_{ijk}(x_i,x_j,x_k)\log\left(\frac{\pi_{ijk}(x_i,x_j,x_k)\pi_k(x_k)}{\pi_{ik}(x_i,x_k)\pi_{jk}(x_j,x_k)}\right)\log\left(\frac{\pi_{ijk}(x_i,x_j,x_k)\pi_j(x_j)}{\pi_{ij}(x_i,x_j)\pi_{kj}(x_k,x_j)}\right)\right]$$

Simulation results show the proposed approximations hold for large sample sizes and odds ratio values not close to independence.

The asymptotic normal approximations presented above can be used to estimate the power of a backwards elimination model selection procedure for selecting the saturated GLL model. In this context we define power of a model selection procedure as the probability of selecting the true model given the specified true model parameters. In the cross-tabulation of three binary variables there are eight cell probabilities that add up to one. Hence, the parameter space is seven dimensional. In the two binary variables case the parameter space has dimension three. Let ξ denote the vector of the chosen parameters, either cell probabilities or combinations of conditional odds ratios and marginal probabilities that uniquely define the contingency table under analysis, depending on the information available. The power of a size α LRT for excluding edge (1,2) from the saturated GLL model with two binary variables can be estimated as

$$P\left[L_{12} > \chi^2_{1;1-\alpha} \;\mid\; \xi\right] \stackrel{a}{=} P\left[Z > \frac{\chi^2_{1;1-\alpha} - AE[L_{12}]}{\sqrt{\text{var}(L_{12})}}\right],$$

where $Z \sim N(0,1)$ and $\chi^2_{1;1-\alpha}$ is the upper α quantile of a chi-squared distribution on one degree of freedom.

In the three binary variables case there are three LRT statistics (L_{12}, L_{13} and L_{23}) for single edge exclusion from the saturated GLL model. The power of selecting the saturated model is the probability that each of these test statistics is greater than $\chi^2_{2;1-\alpha}$, given the values of the chosen parameters in ξ. Power can be approximated with a three-dimensional integral:

$$P\left[min(L_{12}, L_{13}, L_{23}) > \chi^2_{2;1-\alpha} \mid \xi\right] \stackrel{a}{=} \int_D \phi_3\left(\mu,\ \Sigma\right)\ dL_{12}\ dL_{13}\ dL_{23},$$

where $D = \{\chi^2_{2;1-\alpha}, \infty\}^3$ and $\phi_3(\mu, \Sigma)$ is a trivariate normal density with mean vector μ and variance matrix Σ, whose elements are given by the formulae for means, variances and covariances, presented above.
Simulation results were also used to estimate power and to assess the quality of the normal approximations proposed. The main conclusions are that the probability of selecting the saturated model is very sensitive to the total sample size, to the values of the (conditional) odds ratios and to the balance of the contingency tables. The normal approximations to the power of the non signed LRT statistic perform well for large sample sizes and (conditional) odds ratio values not close to one.

4 An Example: University Admissions

Data on graduate admissions to the University of California at Berkeley in 1973, presented by Agresti (2002, page 63), are used to illustrate power calculations. In particular, the associations between admission (A: y or n), gender (G: m or f) and department (D: 3 or 4) is investigated. For these data, $\hat{\psi}_{GA} = 1.02$ and, conditioning on D, $\hat{\psi}_{GA\cdot D=3} = 1.13$ and $\hat{\psi}_{GA\cdot D=4} = 0.92$. For $n_\emptyset = 1710$, the LRT statistic for $H_0 : G \perp\!\!\!\perp A|D$ is 1.05, with a p-value of 0.59, and a backwards elimination model selection procedure chooses model GD, A ($\alpha = 0.05$). Hence, there is no evidence of gender discrimination in the admission process for departments 3 and 4.
To investigate the power associated with this LRT and this model selection procedure, values of $\hat{\psi}_{GA\cdot D=3}$ and $\hat{\psi}_{GA\cdot D=4}$ more extreme than the observed are considered. The five remaining parameters in ξ are selected to be the marginal probability of $D = 3$, $\pi_D(3)$, the probabilities of $G = m$ given $D = d$, $\pi_{G\cdot D}(m, d)$, and the probabilities of $A = y$ given $D = d$, $\pi_{A\cdot D}(y, d)$. These five parameters are set close to their observed values: $\pi_D(3) = 0.54$, $\pi_{G\cdot D}(m, 3) = 0.35$, $\pi_{G\cdot D}(m, 4) = 0.53$ and $\pi_{A\cdot D}(y, 3) = \pi_{A\cdot D}(y, 4) = 0.35$. For the LRT of $H_0 : G \perp\!\!\!\perp A|D$, the power is greater than 0.62 (0.88) if one (both) $\hat{\psi}_{GA\cdot D=3}$ and $\hat{\psi}_{GA\cdot D=4}$ is (are) outside $(0.67, 1.50)$. Hence, a sample of 1710 has enough power to detect a substantively interesting (conditional) association between G and A. For the power of selecting the saturated model the picture is less clear, as can be seen from Table 1. If one of the conditional odds ratios is less than 0.67 and the other is greater than 1.50 then the power is greater than 0.87. However, if they are both less than 0.67 or both greater than 1.50 then the power can be much lower. This is

TABLE 1. *Power of selecting the saturated model for various values of* $\hat{\psi}_{GA\cdot D=3}$ *(in rows) and* $\hat{\psi}_{GA\cdot D=4}$ *(in columns);* $n_\emptyset = 1710$.

	0.25	0.33	0.50	0.67	0.90	1.10	1.50	2.00	3.00	4.00
0.25	0.45	0.49	0.82	0.96	0.99	0.99	0.99	0.99	1.00	1.00
0.33	0.50	0.26	0.49	0.81	0.96	0.99	0.99	0.99	0.99	1.00
0.50	0.83	0.50	0.03	0.16	0.64	0.85	0.98	0.99	0.99	0.99
0.67	0.96	0.82	0.16	0.00	0.10	0.45	0.87	0.98	0.99	0.99
0.90	0.99	0.96	0.65	0.10	0.00	0.00	0.47	0.86	0.99	0.99
1.10	0.99	0.99	0.86	0.46	0.00	0.00	0.11	0.66	0.97	0.99
1.50	0.99	0.99	0.98	0.88	0.48	0.12	0.00	0.17	0.82	0.96
2.00	0.99	0.99	0.99	0.98	0.87	0.68	0.18	0.03	0.50	0.85
3.00	1.00	0.99	0.99	0.99	0.99	0.97	0.83	0.52	0.29	0.54
4.00	1.00	1.00	0.99	0.99	0.99	0.99	0.97	0.86	0.56	0.51

because for such values of $\hat{\psi}_{GA\cdot D}$ and the remaining values of ξ set close to their observed values, the induced conditional association between A and D is small and hence the corresponding edge is not required in the model. The results in Table 1 highlight the need for care when specifying the values in ξ to ensure that power calculations relevant to the hypotheses of interest are being performed.

5 Conclusions

Presented in this paper are methods for estimating the power of single edge exclusion tests and a backwards elimination model selection procedure for a GLL model with two or three binary variables. The methodology presented in this paper can be used for GLL models with four or more binary variables. However, there is currently no straightforward way of generalising the formulae presented, due to the difficulty of handling the parameterisation. In contrast, in the graphical Gaussian framework generalisations are straightforward, as shown by Salgueiro, Smith and McDonald (2003).

Acknowledgments: Portuguese Grant PRAXIS XXI - BD 19873/99.

References

Agresti, A. (2002). *Categorical Data Analysis*, 2nd edition. New Jersey: Wiley.

Salgueiro, M.F. (2002). *Distributions of Test Statistics for Edge Exclusion for Graphical Models*. Ph.D. Thesis, University of Southampton.

Salgueiro, M.F., Smith, P.W.F., and McDonald, J.W. (2003). Power of edge exclusion tests in graphical Gaussian models. *SSRC Methodological Working Paper* ***M03/02***, University of Southampton.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.

On the Analysis of Web Access Logs: Identifying Dense Clusters

Gabriella Schoier[1] and Michael G. Schimek[2]

[1] Dipartimento di Scienze Economiche e Statistiche, Universitá di Trieste, Piazzale Europa 1, I-34127 Trieste, Italy
[2] Institute for Medical Informatics, Statistics and Documentation, Karl-Franzens-University Graz, Engelgasse 13, A-8010 Graz, Austria

Abstract: Web personalization has become an important part of e-commerce. In this paper a solution to the problem of identification of dense clusters in the analysis of Web Access Logs is presented by considering a modification of an algorithm known from social network analysis. The procedure is illustrated by analyzing a portal for children.

Keywords: Web usage mining; Large networks; Blockmodelling, Cluster analysis, Log files.

1 Introduction

The general availability of an ever increasing amount of data coming from the World Wide Web (WWW) is a reality now. Companies which provide their products through the Internet require tools to study and profile their customers in terms of browsing behaviour and personal information. Their aim is gathering useful information and building up business intelligence for the improvement of their web sites and systems (see e.g. Mobasher et. al. (2000), Srivastava et. al. (2000)).
Web Mining is the extraction of interesting and potentially useful patterns and implicit information from artefacts or activity related to the WWW. There are roughly three knowledge discovery domains that pertain to Web Mining: Web Content Mining, Web Structure Mining and Web Usage Mining. The first is the process of extracting knowledge from the content of documents or their descriptions. The second is the process of inferring knowledge from the web organisation and from links between references and referents in the web. The last, also known as Web Log Mining, is the process of extracting interesting patterns in web access logs.
Let us consider a finite set of units (the I.P. addresses) on which one relational variable has been measured (having visited at least m pages in common); this forms a network N (set of units and relation(s) defined over it) (see Wasserman et al (1994)).

In order to analyze such a network one might consider the results developed in two classical social network theories: *the small-world theory* (Kochenet et al (1989)) and *the peer influence theory* (Friedkin (1998)). The first gives evidence that there is a high degree of local clustering in the networks; so an approach for studying the structure of large networks would involve the identification of local clusters and the analysis of the relations within and between clusters. The second theory indicates that, based on an endogenous influence process, close units tend to converge on similar attitudes and thus clusters in a small-world network should be similar along multiple dimensions.
In this paper we present a solution to the problem of identification of dense clusters in the analysis of web access logs, by considering a modification of an algorithm known from social network analysis (see Moody (2001)). The advantage of this approach is a reduced and more flexible structure on which different techniques such as deterministic or stochastic block-modelling , that is a structure which allows us to describe and to interpret a dataset through a block structure. In that way a simplified representation of the existing ties and relations can be obtained (see Schoier (2002)). Moreover a distinctive structure for the degree of similarity within and between clusters is yielded.

2 On the Individualization of Dense Clusters

We exemplify our approach on the log files of the Italian web site `www.girotondo.com`, a portal for children. There are seven different sessions named Bacheca, Corso, Favolando, Giochi, Links, News, Percome, comprising 362 jhtml pages. The period of observation is from the 29/11/2000 to the 18/01/2001. The original file contained 300000 records. Records of log files containing information about any object (with .gif, .jpeg., etc. extension) that is not its internet address are eliminated. In this way we obtain a file indicating the internet address of each visited page. We then proceed with a re-codification of the web pages by transforming their URLs into numbers for easier handling (results in 117 pages). After pre-processing we end up with a file of 10000 records.
The data considered consist of a finite set V of units or vertices (the I.P. addresses) on which one relational variable R (having visited at least $m = 35$ pages in common) is measured. This forms a network N (set of units and relation(s) defined over it).
The network may be represented as a finite graph *G(V,R)* where V represents the set of vertices and R the set of edges composed of pairs of vertices, an actor i is adjacent to actor j if $(v_i, v_j) \in R^2$. The set of all nodes adjacent to node i is the unit's neighbourhood. A path in the network is defined as an alternating sequence of distinct nodes and edges which begin and end with nodes and in which each edge is incident with its preceding and following nodes. Vertex i can reach vertex j if there is a path in the graph

starting with i and ending with j. The length of the path from i to j is given by the number of edges in the path. A network is connected if there is a path of connections between all the pairs of vertices. When the ties are concentrated within subgraphs (that is a graph of which graph vertices and graph edges form subsets of the graph vertices and graph edges of the given graph G) the network is clustered. The level of clustering relates to the fact how uniformly the ties are distributed throughout the network.
A large network requires to individualize local clusters first and then to analyze the internal structures of the clusters or the relations among these clusters. This is exactly what we do in this paper.
Considering our Web data we have a matrix $\mathbf{X}$ (10000×117) which represents one 2-mode network (users $\times$ pages). We refer to Table 1.
A 1-mode network (users $\times$ users) can be obtained via the free program UCINET (Borgatti et al, 1999). The result can be seen in Table 2. The thus obtained matrix has been transformed into a dichotomous matrix of elements 1 denoting cases where the number of visited pages in common exceeds the number 35, and 0 otherwise. This refers to Table 3.
At this point the network is reduced to a set of two position variables collected in the matrix Y (10000×2) using a modified version of the recursive neighbourhood mean algorithm (RNM) proposed by Moody (2001) written in SAS language, the modification consists in the calculation of the mean which is weighting (see step 3). It can be described as follows:
The algorithm of the modified version of RNM is

1. Assign to each I.P. address in the network an uniform random number between 0 and 1 on each of two variables $\mathbf{Y}$
2. Repeat n times
3. Reset each I.P. address value in $\mathbf{Y}$ by weighting according to the number of connections (i.e. the number of shared pages viewed by two users)

This procedure requires in the input, the list of adjacencies, that is the couples of vertices among which exists a relation. See Table 4 At this point the RNM algorithm has been applied. The output of the RNM algorithm is the $\mathbf{Y}$ matrix, displayed in Table 5.
Finally the matrix $\mathbf{Y}$ is used as input for the cluster analysis. On the two positional variables which formed the $\mathbf{Y}$ matrix Ward's minimum variance cluster analysis is carried out. In such a way we obtain a clear clustering that reveals a structure of three clusters (see Moody, 2001, p.268) for the choice of the number of clusters) among the units belonging to the network. These clusters can be used in a faster cluster analysis programs (e.g. Wasserman and Faust, 1994) where techniques like blockmodelling may be applied.

TABLE 1. *Log files matrix.*

I.P. address	PAG. 1	PAG. 2
138.222.202.11	1	0
151.15.169.130	1	0
151.2.15.154	0	0

TABLE 2. *Matrix with the common pages.*

I.P. address	151.20.111.0	151.20.143.184
151.20.111.0	-	37
151.20.143.184	37	-
151.20.9.10	37	37

TABLE 3. *Adjacency matrix.*

I.P. address	151.20.111.0	151.20.143.184
151.20.111.0	-	1
151.20.143.184	1	-
151.20.9.10	1	1

The first cluster, the most numerous one, is formed by the I.P. addresses which have a high frequency of relations, the second one by the I.P. addresses which are not so highly related, and finally the last one representing very few relations. In order to visualize the network the program PAJEK (Batagelj and Mrvar, 2002) has been applied.

3 Conclusions

In this paper we have presented a solution to the problem of identification of dense clusters in the analysis of web access logs, by considering a modification of an algorithm known from social network analysis. Following the cluster analysis eventually block-modelling techniques can be applied. In doing so we have obtained an useful tool to study and profile customers in terms of their browsing behaviour and personal information. This allows us to built up useful business intelligence for the improvement of web sites and the development of systems when data sets are large or even huge.

TABLE 4. *Input of the RNM procedure.*

vertices	vertices
1	3
1	4
1	5
..	..

TABLE 5. *Output of the RNM procedure.*

I.P. address	values of	the Y matrix
151.20.111.0	0.48816	0.42557
151.20.143.184	0.48815	0.42557
151.20.9.10	0.48815	0.42557

References

Batagelj, V. and Mrvar, A. (2002). PAJEK: Program for large Network Analysis.
`http://www.vlado.fmf.uni-lj.si/pub/networks/pajek/`.

Borgatti, S.P., Everett, M. G., and Freeman, L.C. (1999). UCINET for Windows. Software for social network analysis. Analytic Technologies, Harvard, `http://www.analytictech.con`.

Friedkin, N.E. and Johnsen, E. C. (1998). Social position in influence networks. *Social Networks*, **19**, 122–143.

Kochen, M. (1989). *The Small World*. Norwood: Ablex.

Moody, J. (2001). Peer influence groups: Identifying dense clusters in large networks. *Social Networks*, **23**, 261–283.

Schoier, G. (2002). Blockmodelling techniques for web mining. In Härdle, W. and Rönz, B. (eds.) COMPSTAT 2002. Proceedings in Computational Statistics. Physica, Heidelberg, 10–41.

Srivastava, J., Colley, R., Deshpand, M., and Ton P. (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data,
`http://www.maya.cs.depaul.edu/ mobasher/personalization`.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press.

The Method of Dependencies Description with the Help of Optimal Multistage Partitioning

Senko O.V.[1], Kuznetsova A.V.[2], and Kropotov D.A.[1]

[1] Computer Centre of RAS, Vavilova 40,119991, Moscow, Russia
[2] Institute of Biochemical Physics of RAS, Kosygina 4, 119991, Moscow, Russia

Abstract: The method of data analysis is discussed that is based on multistage dichotomic partitioning of regressor variables space.The partitions that provide for best separation of observations with different values of dependent variable are searched inside *apriori* defined families. The problems of statistical verification and optimization of found regularities system are discussed.

Keywords: Data analysis; Partitioning; Validation.

1 Introduction

Suppose that we study dependence of some variable ζ on regressor variables $X_1, ..., X_n$ and our goal is to receive full and valid description of this dependence by related empirical data set. Let vector of regressor variables belongs to some subregion M_x of the multidimensional space $\Re^n$. The various types of dependent variable are admissible. So ζ may be binary variable that is indicator of some class of objects, ζ may be vector of continuous variables and at last ζ may be survival curve. We consider that dependent variable satisfies two conditions. The first one is existence of procedure calculating estimates of ζ mean by related data sets. The estimate of ζ by data set $\tilde{S}$ will be referred to as $\hat{\zeta}(\tilde{S})$. For example $\hat{\zeta}(\tilde{S})$ is the arithmetic mean by all values from $\tilde{S}$ in simple case when ζ is number or $\hat{\zeta}(\tilde{S})$ may be Kaplan-Mayer estimate in case when ζ is survival curve. Let$\hat{\zeta}$ belongs to some set M_ζ . The second condition is existence of distance function ρ that is defined at Cartesian product $M_\zeta \bigotimes M_\zeta$ and has following properties: a)$\rho(\hat{\zeta}', \hat{\zeta}") \geq 0$, b)$\rho(\hat{\zeta}', \hat{\zeta}") = \rho(\hat{\zeta}, \hat{\zeta}')$, c)$\rho(\hat{\zeta}', \hat{\zeta}') = 0$ $\forall \zeta', \zeta" \in M_\zeta$.

2 Multistage Dichotomies

Suppose that we use the empirical data set $\tilde{S}_0 = \{s_1 = (\bar{\zeta}_1, \mathbf{x}_1), ..., s_m = (\bar{\zeta}_m, \mathbf{x}_m)\}$ where $\mathbf{x}_j \in M_x$ and $\bar{\zeta}_j$ is the part of object s_j that is used to calculate the estimate of ζ. For example in case of survival analysis

$\bar{\zeta}_j = (\alpha_j, t_j)$ where t_j is the time of the last observation and α_j indicates if object exists at the time moment t_j. We search the optimal description of dependence as the set $\tilde{Q}$ of subregions from M_x that has the following property. Let $q \in \tilde{Q}$ then it has such neighbor subregion q' that the difference $|\mathbf{M}(\zeta \mid q) - \mathbf{M}(\zeta \mid q')|$ is large enough to be discovered by $\tilde{S}_0$. To construct the optimal set $\tilde{Q}$ we suggest to use multistage partitioning technique. At the first stage the set of optimal dichotomic partitions of M_x is constructed by different independent variables and pairs of variables using families of partitions of different complexity levels. Then the statistical validity of the found dichotomies is evaluated and the first stage output set consisting of l_1 dichotomies is formed: $\tilde{Q}_1 = \{(Q_1, Q_1^c), ..., (Q_{l_1}, Q_{l_1}^c)\}$, where $Q_i \cup Q_i^c$. At the second step optimal dichotomic partitions of subregions $\{Q_1, Q_1^c, ..., Q_{l_1}, Q_{l_1}^c\}$are constructed by correspondent subsets of $\tilde{S}_0 : \{Q_1 \cap \tilde{S}_0, Q_1^c \cap \tilde{S}_0, ..., Q_{l_1} \cap \tilde{S}_0, Q_{l_1}^c \cap \tilde{S}_0\}$. The construction is finished at the k^{th} step when output set of statistically valid dichotomies is empty.

3 Partitions Families

The partition family is defined as the set of partitions with the limited number of elements that are constructed by the same algorithm. The unidimensional and two-dimensional families are considered. The unidimensional families includes partitions of the allowable intervals of single variables. The simplest Family I includes all partitions with two elements (subregions) that are divided by one boundary point. The more complex Family II includes all partitions with no more than three elements that are divided by two boundary points. The two-dimensional families include partitions of two-dimensional allowable areas of pairs of variables. The simplest two-dimensional Family III includes all partitions with no more than four elements that are divided by one boundary point at each axis. At last the most complex Family IV includes no more than nine elements that are divided by no more than two boundary points at each axis. The use of partitions families with several levels of complexity has the following explanation. Some more complicated regularities cannot be discovered with the help of simple partitions. On the other hand the use of complex partitions families when regularities are actually simple leads to distortions of boundary points and to decrease of statistical validity due to overfitting effects. The use of more complicated models from families II-IV leads to arising of great variety of statistically valid regularities that actually are induced by more simple ones. We suggest to eliminate superfluous regularities by eliminating from output set models that are reduced to several more simple dichotomies and at least one of these dichotomies has the same or greater level of statistical validity.

4 Optimal Dichotomic Partitions

An optimal dichotomic partition may be constructed by single variable using unidimensional partitions families or by pair of variables using 2-dimensional partitions families. Suppose that some partition R consists of subregions $q_1, ..., q_r$ where $\bigcup_{i=1}^{r} q_r = M_x$. The partition induces the partition of data set on subsets $\tilde{S}_1, ..., \tilde{S}_r$, where subset $\tilde{S}_i$includes all objects from $\tilde{S}_0$ with the vector of independent variables belonging to q_i . The value of partition quality functional for is calculated as $F(\tilde{S}_0, R) = \max_{i \in \{1,...,r\}} \{\rho[\hat{\zeta}(\tilde{S}_i), (\tilde{S}_0)]\}$, where $m_i = \mid \tilde{S}_i \mid$. Suppose that the maximal value of partition quality functional is achieved on partition $R_0 = \{q_1^0, ..., q_r^0\}$ and index i equal i_0. Then the optimal dichotomic partition of subregion M_x is formed as pair of subregions $\{q_{i_0}^0, M_x \setminus q_{i_0}^0\}$.

5 Validation

The main problem in discussed partitioning technique is statistical validation of discovered regularities. The simple solution exists when the initial data set is large enough to form besides training set $\tilde{S}_0$ also the control set $\tilde{S}_c$. You can find the optimal partition by $\tilde{S}_0$ and to estimate statistical validity using $\tilde{S}_c$ with the help of standard statistical tests. But we cannot use the same data set for the search of optimal partitions and for the estimation of statistical validity. So in case of relatively small size of initial data set we suggest to use the technique based on permutations to test the null hypothesis $\mathbf{H}_0$ about independence of ζ on variables $X_1, ..., X_n$. The probability that the data set may casually arise when $\mathbf{H}_0$ is true with the quality functional meaning at optimal partition exceeding such optimal functional meaning for true data set may be used as the measure of statistical validity (p-value) of regularity discovered by . Let $\tilde{f} = \{f_1, ..., f_m\}$ is permutation of numbers from the set $\{1, ..., m\}$. The artificial training set may be constructed from $\tilde{S}_0$ as $\tilde{S}_f = \{s_1 = (\bar{\zeta}_{f_1}, \mathbf{x}_1), ...s_m = (\bar{\zeta}_{f_m}, \mathbf{x}_m)\}$. In case when $\mathbf{H}_0$ is true and objects to data sets are selected from the same distribution and independently the probability of $\tilde{S}_0$ coincides with the probability of $\tilde{S}_f$. So to estimate the p-value it is sufficient to calculate the ratio of permutations that allow receiving optimal quality functional value exceeding optimal value at $\tilde{S}_0$ to the full number of permutations.

6 Realization

The main drawback of suggested approach is too great amount of calculating that is necessary to find all optimal partitions and to estimate their statistical validity especially when permutation test is used. Experiments have shown that the variant of discussed approach based on interactive

mode may be successfully implemented even at common PC with the Pentium type single processor. Such realization of the approach includes the search of all optimal dichotomies using partitions families I-III. The user selects the most interesting dichotomies from the found set and estimates their statistical validity with the help of permutation test or some other technique if such possibility exists. Significantly more complete version of the approach is realized at the parallel system. The use of parallel computing allows to find all dichotomies using partitions families I-IV together with evaluation of their statistical validity.As it was noted previously in section **Partitions families** the use of complicated models leads to arising to too great number of revealed regularities. The following approach to eliminating of superfluous regularities was suggested. The significance levels of complicated models from families II-IV are compared with significance of more simple regularities that are found for the same variables. We consider that regularity is reduced to more simple one and exclude it from the output list if its significance level is worse than significance of at least one simple regularity.

7 Experiments

The developed approach was successfully used in several tasks of medical data analysis (Senko et al, 2001). However the correct evaluation of its effectiveness is possible in case when true probability distributions are known. So the Monte-Carlo simulation was used to test the performance of optimal partitioning technique. We consider the scenario when dependent variable ζ belongs to the set $\{0, 1\}$ with equal probabilities of 0 and 1. Regressor variables $X_1, ..., X_n$ are independent and are distributed uniformly at cut [0,1], when $\zeta = 1$. In case when $\zeta = 0$ some of regressor variables are distributed uniformly inside subregions of multidimensional cube $[0, 1]^n$ that can be described by dichotomic partitions from models I-III. The 10 artificial data sets with $m = 100$ observations $n = 10$ regressor variables were received subsequently with the help of the same random numbers generator.Variables $X_1 - X_8$ were mutually independent. Besides variables $X_1 - X_5$ were distributed uniformly at cut[0,1] for both values of ζ. So ζ did not depend on variables from this group. Probability distributions of variables X_6 and X_7 were described by unidimensional models with one boundary point corresponding to family I: $\mathbf{P}(X_6 \in [0, 0.5] \mid \zeta = 0) = 0.95$, $\mathbf{P}(X_6 \in (0.5, 1.0] \mid \zeta = 0) = 0.05$, $\mathbf{P}(X_6 \in [0, 0.3] \mid \zeta = 0) = 0.7$, $\mathbf{P}(X_6 \in (0.3, 1.0] \mid \zeta = 0) = 0.3$. Probability distribution of variables X_8 were described by unidimensional model with two boundary points: $\mathbf{P}(X_8 \in [0.3, 0.7] \mid \zeta = 0) = 0.75$, $\mathbf{P}(X_8 < 0.3 \vee X_8 > 0.7 \mid \zeta = 0) = 0.25$. Joint distribution of variables X_9 and X_{10} was described by 2-dimensional model with one boundary points for each variable: $\mathbf{P}(X_9 < 0.4 \& X_1 0 < 0.7 \mid \zeta = 0) = 0.6$,

$\mathbf{P}(X_9 > 0.4 \vee X_{10} > 0.7 \mid \zeta = 0) = 0.4$. The statistical validity of revealed regularities was evaluated with the help of permutation test that is described in section **Validation**. Only regularities with validity estimated at the significance level $p < 0.01$ were considered. The unidimensional one boundary regularity related to variable X_6 was revealed by selection of optimal model from family I for all 10 generated data sets. The same regularity related to variable X_7 was revealed for 8 data sets only. The unidimensional regularity with two boundary point related to variable X_8 was revealed by selection of optimal model from family II for 6 generated data sets.Only two regularities from 10 were included to output list of 2-dimensional regularities related to variable X_9 and X_{10}. These regularities were revealed by selection of optimal model from family III. The cause of the poor results in case of 2-dimensional regularities may be the relatively low difference (0.6/0.4) of ζ conditional means in subregions of regressor variables space. The only 3 regularities related to variables from group $X_1 - X_5$ were selected. The two of them were unidimensional with two boundary points and one was 2-dimensional with one boundary at each variable. The experiments demonstrated that approach allows to reveal regularities but large scale investigations are necessary to evaluate the sets of data analysis tasks where it is effective.

Acknowledgments: Special thanks to RFFI (grant 03-01-00580)

References

Chou, P. (1991). Optimal partitioning for classification and regression trees. *IEEE IEEE Transactions on Pattern Analysis and Machine Intelligence*, **2**, 340–354.

Senko, O.V. and Kuznetsova, A.V. (1999). *The use of partitioning for analysis of biomedical data*. Proceedings of 14th International Workshop on Statistical Modelling, Graz, Austria, pp. 656-659.

Senko, O.V. and Kuznetsova, A.V., Matchak, G.N., Vakhotsky, V.V., Zabotina, T.N., and Korotkova, O.V. (2000). The prognosis of survivance in solid tumor patients based on partitioning of immunological parameters ranges.*Journal of Theoretical Medicine*, **2**, 317–327.

Senko, O.V., Kuznetsova, A.V. (1998). The use of partitions construction for stochastic dependencies approximation.*Proceedings of the International Conference on Systems and Signals in Intelligent Technologies*. 28-29 September, Minsk (Belarus), 291-297.

A Hierarchical Bayesian Approach for the Evaluation of Surrogate Endpoints in Multiple Randomized Clinical Trials

Ziv Shkedy[1], Franz Torres[1], Tomasz Burzykowski[1], and Geert Molenberghs[1]

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

Abstract: The issue of surrogate endpoints in randomized clinical trials arises whenever the time needed to observe the primary endpoint is very long or if the primary endpoint is very expensive to observe. In theses cases, one may assess the treatment effect on a surrogate endpoint instead on the primary endpoint and reduce the duration or price of the trial. In this paper we use hierarchical Bayesian models to evaluate a potential surrogate endpoint in multiple randomized clinical trials. The methods are illustrated using data from three randomized clinical trials.

Keywords: Hierarchical Bayesian models; Individual level surrogacy; Surrogate endpoints; Trial level surrogacy.

In randomized clinical trials, the main interest is to assess the treatment (Z) effect on the primary endpoint (T). However, in some cases, the time needed to observe the endpoint of interest can be long (for example, if the primary endpoint is time to event) or very expensive. In these cases, one might benefit from using a surrogate endpoint(S), that would allow to determine the treatment effect quicker or in a less expensive way.
In his landmark paper, Prentice (1989) proposed a formal definition of a surrogate endpoint and suggested operational criteria for its validation in the case of a single trial and single surrogate. According to the definition, a surrogate endpoint is a variable for which a test of the null hypothesis of no treatment effect is also a valid test of the corresponding null hypothesis for the true endpoint. In view of some limitations of Prentice's criteria, Freedman, Graubard, and Schatzkin (1992) proposed to use the proportion of treatment effect explained by the surrogate endpoint as a measure of the validity of a potential surrogate. Several authors have pointed towards drawbacks of the measure. For instance, De Gruttola *et al.* (1997) and Buyse and Molenberghs (1998) have shown that the proportion of treatment effect explained by the surrogate is not truly a proportion, since it is not restricted to the $[0, 1]$ interval. As an alternative, Buyse and Molenberghs (1998) proposed to replace the proportion of treatment effect explained by

the surrogate by two measures closely related to it: the relative effect and the adjusted association. The first one, defined at the population level, is the ratio of the overall treatment effect on the true endpoint over that on the surrogate endpoint. The second one is the individual-level association between both endpoints, after accounting for the effect of treatment.
In this paper we focus on the meta-analytic approach, that is, the situation when a potential surrogate is evaluated using data from multiple trials. We further assume that the distribution of the true and surrogate endpoint come from the exponential family and that true treatment effects on the endpoints are given by

$$\begin{array}{l} g\{E(S|Z=1)\} - g\{E(S|Z=0)\} = \alpha, \\ g\{E(T|Z=1)\} - g\{E(T|Z=0)\} = \beta, \end{array} \tag{1}$$

where $g()$ denotes an appropriate link function. Within the meta-analytic approach the first goal is to establish the association between β and α to asses the quality of the surrogate at the trial level. To this aim, the precision of the prediction of the treatment effect on the true endpoint form the effect on the surrogate, should be assessed. This can be achieved by formulating a model for the joint distribution of treatment effects $[\alpha, \beta]$, or a model of the conditional distribution $[\beta|\alpha]$. Note that a joint model $[\alpha, \beta]$ imposes a conditional model for $[\beta|\alpha]$ but one can specify a model for $[\beta|\alpha]$ without specifying the joint model. The second goal is to assess the quality of individual level surrogacy, i.e., the precision of the prediction of the true endpoint from the surrogate for an individual patient. This can be evaluated considering the association between the two endpoints in the joint distribution o f! S and T given Z, $[T, S, |Z]$.
The evaluation of a surrogate endpoint within the meta-analytic setting has been discussed, e.g., by Daniels and Hughes (1997) and Buyse et al (2000). Both papers considered a multiple trial setting with normally distributed true and surrogate endpoints and proposed a two-stage model for the evaluation of the potential surrogate. Daniels and Hughes (1997) assumed that only summary data from the trials were available. They used a hierarchical Bayesian model for the estimated treatment effects $[\hat{\alpha}, \hat{\beta}]$, in which the joint distribution of the estimated effects was specified at the first stage and the conditional distribution of $[\beta|\alpha]$ was specified in the second stage. Buyse et al (2000) assumed the availability of individual-patient data and formulated a two-stage model, with the joint distribution $[T, S|Z]$ specified at the first stage and the joint distribution of the treatment effects $[\beta, \alpha]$ specified at the second s! tage. The advantage of the model proposed by Daniels and Hughes (1997) is that one does not need to specify the joint distribution of T and S. However, the price for this advantage is that the quality of the individual level surrogacy cannot be assessed, what is possible in the approach developed by Buyse et al (2000). In this paper we consider the Bayesian approach under the assumption that individual data are available.

We consider the following linear predictors for T and S

$$\begin{cases} E(S_{ij}|Z_{ij}) = \mu_{Si} + \alpha_i Z_{ij}, \\ E(T_{ij}|Z_{ij}) = \mu_{Ti} + \beta_i Z_{ij}. \end{cases} \tag{2}$$

Here α_i and β_i are trial-specific treatment effects, μ_{S_i} and μ_{Ti} are trial-specific intercepts and S_{ij} and T_{ij} are the surrogate and the true endpoints, respectively, for subject j $(j = 1, 2, \ldots, n_i)$ in trial i $(i = i, 2, \ldots, N)$. We further assume that the two endpoints are normally distributed. Thus, at the first stage of the hierarchical model we specify the following joint distribution of T_{ij} and S_{ij}:

$$\begin{pmatrix} S_{ij} \\ T_{ij} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_S + m_{Si} + (\alpha + a_i)Z_{ij} \\ \mu_T + m_{Ti} + (\beta + b_i)Z_{ij} \end{pmatrix}, \Sigma \right), \tag{3}$$

where Σ is given by

$$\Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ \sigma_{ST} & \sigma_{TT} \end{pmatrix}. \tag{4}$$

At the second stage of the model the priors for the 'fixed' effects are specified:

$$\begin{aligned} &\mu_S \sim \mathrm{N}(0, \theta^2_{\mu_S}), \\ &\mu_T \sim \mathrm{N}(0, \theta^2_{\mu_T}), \\ &\alpha \sim \mathrm{N}(0, \tau^2_{\alpha}), \\ &\beta \sim \mathrm{N}(0, \tau^2_{\alpha}). \end{aligned} \tag{5}$$

For the precision parameters in (5) (flat) hyperprior models were specified using Gamma distributions, e.g., $\theta^{-2}_{\mu_S} \sim \mathrm{gamma}(0.001, 0.001)$, etc. Similar to the model proposed by Daniels and Hughes (1997) we need to specify a prior distribution to model the association between the treatment effects of the two endpoints. Note that, while Daniels and Hughes (1997) based their model on $[\beta|\alpha]$, Buyse *et al.* (2000) used the joint distribution of the random effects $[m_{Si}, m_{Ti}, a_i, b_i]$ in order to evaluate trial level surrogacy. In the current model we follow the latter approach and specify the prior model for the joint distribution of the random effects to be

$$\begin{pmatrix} m_{Si} \\ m_{Ti} \\ a_i \\ b_i \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, D \right), \quad D = \left(\begin{array}{cc|cc} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ d_{ST} & d_{TT} & d_{Ta} & d_{Tb} \\ \hline d_{Sa} & d_{Ta} & d_{aa} & d_{ab} \\ d_{Sb} & d_{Tb} & d_{ab} & d_{bb} \end{array} \right). \tag{6}$$

As the hyperprior distribution for the covariance matrices in (3) and (6), a Wishart distribution is assumed:

$$D^{-1} \sim \mathrm{Wishart}(R_D) \quad \text{and} \quad \Sigma^{-1} \sim \mathrm{Wishart}(R_\Sigma). \tag{7}$$

In order to assess trial-level surrogacy, Buyse *et al.* (2000) proposed to use the coefficient of determination deifned as:

$$R^2_{trial(f)} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}. \tag{8}$$

Similarly, to measure individual-level surrogacy Buyse *et al.* (2000) proposed to use the coefficient of determination given by

$$R^2_{indiv} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}}. \tag{9}$$

Indeed, $R^2_{trial(f)} = 1$ and $R^2_{indiv} = 1$ indicate perfect surrogacy at trial and individual level, respectively.

To avoid computational problems, Buyse et al (2000) proposed a reduced model in which the linear predictors of S and T do not include trial-specific intercepts. In the hierarchical model, the likelihood at the first stage of the model can be specified by omitting the trial specific random intercepts from (3). This leads to specify that

$$\begin{pmatrix} S_{ij} \\ T_{ij} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_S + (\alpha + a_i)Z_{ij} \\ \mu_T + (\beta + b_i)Z_{ij} \end{pmatrix}, \Sigma \right), \tag{10}$$

At the second stage of the model, the prior distribution the random effects, (a_i, b_i), was assumed to be bivariate normal with mean 0 and covariance matrix D. Note that the covariance matrix D is the 2×2 right bottom sub matrix in (6) and is assumed to follow a Wishart distribution, $D^{-1} \sim$ Wishart(R_D). Other prior and hyperprior models remain the same as in the full model. For the reduced model the coefficient of determination (8), measuring the trial-level surrogacy, reduces to $R^2_{trial(r)} = d^2_{ab}/d_{aa}d_{bb}$.
For illustration, we consider data from four randomized multicenter trials in advanced ovarian cancer, previously analyzed by Buyse et al (2000). The data were collected for a purpose of a meta-analysis considering the comparison of cyclosphosphamide plus cisplatin with cyclosphosphamide plus adriamycin plus cisplatin (Ovarian Cancer Meta-Analysis Project, 1991). The true endpoint T_{ij} is defined as Log(survival time in years) and the surrogate endpoint S_{ij} is taken as Log(progression-free survival time in years). We used center as the unit of analysis given that the number of trials is insufficient to applied meta-analytic methods. A total of 50 centers were available for the analysis, with the number of patients varying 2 to 274 per center.
We fitted the hierarchical Bayesian models (3)–(7) and (10) using MCMC simulation with 9000 iteration following a burn-in period of 1000 iterations. Table 1 presents the maximum likelihood estimates for both $R^2_{trial(f)}$ and

TABLE 1. *$R^2_{trial(f)}$ and R^2_{indiv}. The full fixed effects model corresponds to the model in Eq. (2), while the reduced fixed effects model corresponds to the model in Eq. (2) without trial-specific intercepts. The results for the fixed effects model were obtained by Buyse et al (2000). $R^2_{trial} \equiv R^2_{trial(f)}$ and $R^2_{trial} \equiv R^2_{trial(r)}$ for the full and the reduced model, respectively.*

MODEL	Trial level R^2_{trial}	Individual level R^2_{indiv}
Full (Fixed)	0.940 (0.017)	0.886 (0.0006)
Full(Bayesian)	0.938 (0.038)	0.885 (0.0006)
Reduced (Fixed)	0.928 (0.020)	0.888 (0.0006)
Reduced (Bayesian)	0.925 (0.048)	0.885 (0.0006)

R^2_{indiv} obtained from the fixed effects model and reported in Buyse et al (2000), as well as the posterior means obtained from the corresponding Bayesian model. The point estimates are comparable, while the standard errors of $R^2_{trial(f)}$ are greater for the hierarchical Bayesian model. This results in credible intervals for the posterior means which are wider than the confidence intervals for the ML estimates.

The results presented in this work suggest that the use of the hierarchical Bayesian modelling for the meta-analytic approach to the validation of surrogate endpoints is completely feasible.

References

Buyse, M. and Molenberghs, G. (1998). The validation of surrogate endpoints in randomized experiments. *Biometrics,* **54**, 186–201.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics,* **1**, 49–67.

Daniels, M.J. and Hughes, M.D. (1997). Meta-analysis for the evaluation of potential surrogate markers, *Statistics in Medicine,* **16**, 1965-1982.

De Gruttola, V., Fleming, T.R., Lin, D.Y., and Coombs, R. (1997). Validating surrogate markers - are we being naive? *Journal of Infectious Diseases,* **175**, 237–246.

Freedman, L.S., Graubard, B.I., and Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine,* **11**, 167–178.

Prentice, R.L. (1989). Surrogate endpoints in clinical trials: definitions and operational criteria. *Statistics in Medicine,* **8**, 431–440.

Customer Data Mining with Clustering Technique

Hizir Sofyan[1] and Jianqiu Wang[1]

[1] Institut für Statistik und Ökonometrie, HumboldtUniversität zu Berlin, Spandauer Str. 1, 10178 Berlin, Germany, Phone: +49-30-2093-5623, Fax : +49-30-2093-5959, Email: hizir@wiwi.hu-berlin.de

Abstract:
Customer analysis is the study of customers and their behaviors. Nowadays, data mining is a frequently adopted technique to conduct customer analysis, so that the companies can discover valuable information among their customer data, such as the segmentation of the customers. Clustering is one of data mining techniques which supports the customer segmentation. It partitions the observations into disjoint groups such that the profiles of objects in the same groups are relatively homogenous, whereas the profiles of objects in different groups are relatively heterogeneous.
This paper presents New Condorcet Criterion, a non-hierarchical clustering technique. This technique is particularly work well for a data set that consists of categorical data. We apply the method to analyze the customer behaviors of `XploRe` statistical software.

Keywords: Customer analysis; Data mining; Clustering technique.

1 Introduction

The sensational developments in the field of information technology in the last ten years have eased a lot of complications that were related to the collection of bulky data. It has created the necessity for automated information discovery from data, which has lead to a growth of the promising field called data mining (Fayyad et al, 1996).
Data mining tries to discover patterns and relationships hidden in the data using suitable statistical models and techniques (Chen, Han and Yu, 1996). Therefore, data mining may yield profitable results for almost every organization that collects data on its customers, markets, products or processes. Clustering is considered as one of data mining tools. According to the research held by KDD Nuggets (`http://www.kdnuggets.com`), clustering is the most frequent method applied in data mining. Clustering is aimed to discover a group and to identify interesting distributions and patterns within the data.

This paper presents New Condorcet Criterion, a non-hierarchical clustering method. This method is particularly work well for a data set that consists of categorical data. We apply the method to analyze the customer behaviors of `XploRe` statistical software. Based on the characteristics of the cluster members, we outline how the result of this analysis may be used for marketing strategy of `XploRe`.
The rest of the paper is organized as follows. The next section reviews the methodology of Condorcet Criterion. The third section presents the data and mining technology. In the fourth part, the result and discussions are presented. Final section contains some concluding remarks.

2 Condorcet Criterion

The clustering algorithm used in this paper is based on the New Condorcet Criterion (NCC) of Michaud (1997). It is inspired by Condorcet (1743-1794)'s work on finding a desirable way to aggregate votes (rankings) in an election (Michaud, 87).
The NCC is defined for categorical attributes. The distance between attribute values as 1 if two elements have different values and 0 otherwise. The distance between two elements can be viewed as a modified hamming distance, that is, the number of attributes for which the two elements i and j is the number of "judges" who "disagree" about whether elements i and j should be in the same class (and $m - d_{ij}$ is the number of agreements).
The NCC combines intraclass agreement as well as interclass disagreement such that "good" partitions, i.e. those with small intraclass distances and large interclass distances, get higher values of the criterion function (Grabmeier and Rudolph, 2002).

3 The Data and Mining Technology

The aim of our analysis is to identify a number of clusters of users who have downloaded the statistical software `XploRe` by performing a cluster analysis of its download profiles, which could function as the base for the development of the marketing strategy.
The collected raw data of `XploRe` user consists of 2593 profiles of individuals who have downloaded the statistical software `XploRe` from October 11, 2001 to March 13, 2003. A free trial version of `XploRe` can be downloaded from the homepage at `http://www.xplore-stat.de`.
Before the downloading, users are asked to participate an online survey. The online questionnaire composes mainly two parts. All questions (except for email address) are answered by selecting items from possible responses. We have made some improvements in the survey comparing the previous practice, which was conducted by Sofyan and Werwatz (2001). The new variables concerning the benefit sought by the users and features of

their research were added to the questionnaire. In addition, the values of some some variables were also reformed. We choose IBM's Intelligent Miner (`http://www.software.ibm.com/data/ iminer`) for our analysis because it employs condorcet criterion which is particularly well-suited for categorical data sets.

4 Results and Discusions

The prior analysis (Sofyan and Werwatz, 2001) has indicated that the optimal partition of the data is not necessary to be the one, which is with highest statistical goodness value but has no meaningful characteristics for marketing. Bearing the segmentation requirements in mind, the chosen partition should have relative high goodness value of statistics and, at the same time, could deliver a handful groups that can be handled and targeted by the marketer. The targeted groups should be within the reach of and sensitive to the marketing instruments. Therefore, one must carefully consider about the clustering segments, whether the partition could be used to develop the marketing strategy for the target market.
The final segmentation had five variables and four clusters. The five variables are *work field, where work, resource of first learn,* `XploRe` *version* and *OS platform*. With these five variables, the four cluster segmentation achieves relative high of NCC value (0.6002) and good interpretation of the data comparing the other segmentation. As mentioned before, the final chosen segmentation should not only achieve the high statistic value, but also could deliver a rational description of the data. Therefore, we dropped the results that have lower NCC values and the results which are with high NCC value but difficult to tape meaningful characteristics for the customer groups.
Figure 1 shows the visual result of the clustering. The character of cluster Cluster 1 is dominant by the value "internet" of variable *first learn*. Therefore, the segment of Cluster 1 called as **Internet surfer**. This group of users are more like to look for information through internet. They download the local version of `XploRe`, use windows as platform. They work in widespread field. The working places of them are similar distributed as whole sample, mainly in university and home.
Cluster 2 and Cluster 3 are determined by two dominant variables, *first learn* and *where work*. Users from Cluster 2 work mainly at "university (88%)" and the main information resource is "friends (39%)". Users from Cluster 3 work mainly at "home (67%)" and they first get learn `XploRe` through "some unidentified resources (44%)". Thus, Cluster 2 is **Academia**, who work at university and Cluster 3 is **Home worker**, who work at home. **Academia** and **Home workers** also mainly download local version of `XploRe` and use windows as platform. But the **Academia** work mainly in the field of econometrics, while the **Home workers** work in

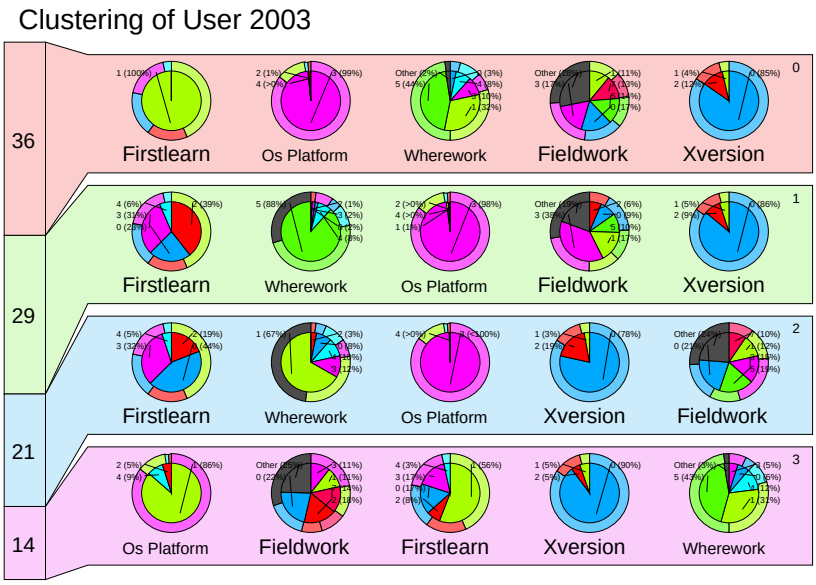


FIGURE 1. *Visualization of the optimal partition.*

widely spread fields with finance and actuarial science in a relative high percentage 17.

`XploRe` users from Cluster 4 are **Linux user**, who are indicated by the variable "platform". 86% of them use Linux as platform. **Linux user** prefer to download the local version of `XploRe` as well. They work in widely spread fields and places, get information from different resources, among which "Internet" has a relative dominant position (56%).

5 Conclusion

In this paper, we have presented the results of a cluster analysis of 2593 profiles of individuals who have downloaded the statistical software `XploRe`. Each profile consisted of a set of variables that are the responses to a mandatory online questionnaire preceeding the actual downloading process. Using New Condorcet Method particularly suited for our categorical data, we arrived at a partition consisting of four clusters: **Internet surfer, Academia, Linux user** and **Home worker**. All the groups except Linux users work with windows systems and engage in economic study of finance or econometrics. **Internet surfers** are characterized by their dominant usage of Internet as the information resources. **Academia** are researchers from university with more academic background. They heavily depend on

personal communication channels. **Home workers** is a mix group. They work mainly at home, get information from various resources, among which Publications/Journal has a relative important position. The **Linux users** are sophisticated computer users who work under Linux, have natural science background and make relative heavy use of the internet.

Acknowledgments: Special Thanks to Prof. Dr. Wolfgang Härdle of Institut für Statistik und Ökonometrie for his useful comments suggestions.

References

Chen, M., Han, J., and Yu, P.S. (1996). Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, **8**(6), 866–883.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996) *Advances in Knowledge Discovery and Data Mining*, The AAAI Press, Menlo Park, CA.

Grabmeier, J. and Rudolph, A. (2002). Techniques of cluster algorithms in data mining. *Data Mining and Knowledge Discovery*, **6**(4), 303–360.

Härdle, W., Klinke, S., and Müller, M. (1999). *XploRe Learning Guide*. Heidelberg: Springer-Verlag.

Michaud, P. (1987). Condorcet: A man of the avant-garde. *Applied Stochastic Models and Data Analysis*, **3**, 173–189.

Michaud, P. (1997). Clustering techniques. *Future Generation Computer Systems*, **13**, 135–147.

Sofyan, H. and Werwatz, A. (2001). Analyzing XploRe download profiles with intelligent miner.*Computational Statistics*, **16**, 645–479.

Multivariate Plackett-Dale Inference to Study the Inheritance of Longevity in a Belgian Village

Fabián Tibaldi[1], Bart Van de Putte[2], Helena Geys[1], Geert Molenberghs[1], Koen Matthijs[2], and Robert Vlietinck[3]

[1] Center for Statistics Limburgs Universitair Centrum, Universitaire Campus, Building D, B 3590 Diepenbeek, Belgium, email:fabian.tibaldi@luc.ac.be
[2] Center for Population and Family Studies, Department of Sociology, Katholieke Universiteit Leuven, Belgium
[3] University Maastricht, Division of Genetics and Molecular Cell Biology Universiteitssingel 50, 6229 ER Maastricht, The Netherlands Center of Human Genetics, Katholieke Universiteit Leuven, Belgium Kapucijnenvoer 33 Blok J, 3000 Leuven, Belgium

Abstract: We consider a Plackett-Dale model to study familial transmittance of longevity. We focus the analysis on associations between mother, father and first child, and therefore we work with family clusters of equal size. We propose a series of tests to perform inferences on the model parameters. The methodology is applied to a demographic database of a Flemish village (18^{th}-20^{th} century). The main substantial conclusion is that familial transmission happens mainly via the mother. We explore the impact of such other factors as censoring, gender effect, age at death, etc. This paper complements the results of Matthijs et al (2002) and suggests further analyses to better understand the precise mechanisms behind these associations.

Keywords: Plackett-Dale model; Multivariate survival model; Pseudo-likelihood inference; Familial clustering; Correlation.

1 Introduction

The main topic of this work is to propose a number of inferential tools to test the parameters of a multivariate marginal survival model. We explain how the methodology works and we apply it to the study of associations between longevity of family members in a small Flemish village. Each family is treated as a cluster and we will use a multivariate Dale model for survival data combined with pseudo-likelihood ideas. The main substantial topic to be addressed are differences in the influences of fathers and mothers on the female offspring's longevity.
Moerzeke is a small village in the center of Flanders, the Dutch speaking part of Belgium, within the province of East Flanders. It is a geographical

isolate as it is almost completely surrounded by the river Scheldt.
The information in the Moerzeke database is drawn from church and civil registers. In Belgium, these sources are of good quality and appropriate for populations studies. The database contains all individuals who were born, married or died in Moerzeke.

2 Statistical Model

Pseudo-likelihood ideas are used to estimate the parameters and a number of inferential tools are proposed. We consider the survival times T_j of mother, father, and first child ($j = 1, 2, 3$) of 457 families with complete information on dates of death and we observe a vector of covariates $\boldsymbol{Z}$. Marginal Weibull distributions for each survival time are assumed. Let us consider the individual information of family i expressed in vector format as $(T_{i1}, T_{i2}, T_{i3}, \boldsymbol{\Delta}_{i1}, \boldsymbol{\Delta}_{i2}, \boldsymbol{\Delta}_{i3}, z_{i1}, \ldots, z_{in_3})$ so that $\boldsymbol{W}_{ij} = (\boldsymbol{T}_i, \boldsymbol{\Delta}_i, \boldsymbol{Z}_i)$ are the values for a particular cluster i and survival time j within cluster. The $\boldsymbol{\Delta}_{ij}$ variable indicates whether the lifetimes is observed or not.
The pseudo-likelihood function to estimate the parameters of this model is constructed along the lines of Le Cessie et al (1994) and Renard et al (2002) by considering all three possible pairs of outcomes on an individual $(\boldsymbol{W}_{1r}, \boldsymbol{W}_{2\ell})$ $(\boldsymbol{W}_{1r}, \boldsymbol{W}_{3\ell})$ and $(\boldsymbol{W}_{2r}, \boldsymbol{W}_{3\ell})$. Those pairs produce $f_{T_rT_\ell}(\boldsymbol{W}_{ir}, \boldsymbol{W}_{i\ell})$ with $r < \ell$, $r = 1, 2, 3$ and $\ell = 1, 2, 3$, where $f_{T_rT_\ell}$ is the density function of the Plackett-Dale distribution (Dale 1986 and Mardia 1970). In this case the dependency can be defined using a *global cross-ratio* at (t_r, t_ℓ) given by $\theta_{r\ell}(t_r, t_\ell)$. The Plackett distribution is obtained for constant cross-ratio $\theta_{r\ell}(t_r, t_\ell) \equiv \theta$ (Plackett 1965, Mardia 1970).
We can define then

$$\ln p\ell(\boldsymbol{\Phi}) = \sum_{i=1}^{N} \sum_{(s,t)\in S} \ln f_{T_sT_t}(\boldsymbol{W}_{is}, \boldsymbol{W}_{it}, \boldsymbol{\Phi}), \tag{1}$$

where S is the set of all possible pairs of outcomes of interest and $\boldsymbol{\Phi}$ the vector of parameters.
The pseudo-likelihood estimator $\widehat{\boldsymbol{\Phi}}$ is defined as the maximiser of (1). Consistency has been shown by Arnold and Strauss (1991), le Cessie and van Houwelingen (1994), and Geys, Molenberghs, and Ryan (1999). The parameters of this model and their standard errors can be estimated by means of the maximum likelihood method and the asymptotic normality results provide an easy way to consistently estimate the asymptotic covariance matrix. Precisely, $\widehat{\boldsymbol{\Phi}}$ converges in probability to the true parameter value $\boldsymbol{\Phi}_0$, and $\sqrt{N}(\widehat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}_0)$ converges in distribution to $N_q(\boldsymbol{0}, J(\boldsymbol{\Phi}_0)^{-1}K(\boldsymbol{\Phi}_0)J(\boldsymbol{\Phi}_0)^{-1})$ with $J(\boldsymbol{\Phi})$ and $K(\boldsymbol{\Phi})$ appropriate defined.
This asymptotic normality result provides an easy way to consistently estimate the asymptotic covariance matrix. A further advantage of the PL

approach is the close connection of pseudo-likelihood with likelihood, enabling one to construct pseudo-likelihood ratio and pseudo-score test statistics that have easy-to-compute expressions and intuitively appealing distributions (Aerts et al 2002).

3 Test Statistics

To test the parameters of the model several tools can be used as Wald, score or likelihood ratio tests. However, while point estimation and asymptotic normality have already been established, we need to construct the pseudo-likelihood counterparts to classical inferential tools such as ratio test statistics and score test statistics. Particularly, to perform a test for the association parameters of the model, we need to extend the Wald, score and likelihood ratio test statistics to the pseudo-likelihood framework. It is important to note that the strategies proposed are not restricted to those parameters and it can be applied to any other model parameter.
Association parameters θ_{ij} equaling one indicate independence between T_i and T_j. This can be translated in terms of hypotheses such as

$$\mathrm{H}_0 : \theta_{r\ell} = 1 \qquad \theta_{r\ell} \in I\!R_{\geq 0} \qquad r, \ell = 1, 2, 3.$$

More generally, let us assume we are interested in an hypothesis of the type $\mathrm{H}_0 : \varphi = \varphi_0$ where φ denotes a q-dimensional subvector of the p-dimensional vector of regression parameters $\mathbf{\Phi}$ and write $\mathbf{\Phi} = (\varphi', \beta')'$.
To construct the Wald test we use the asymptotic normality properties of the pseudo-likelihood estimators. We use the following result

$$W^* = N(\widehat{\varphi} - \varphi_0)' \Sigma^{-1}_{\varphi\varphi} (\widehat{\varphi} - \varphi_0) \sim \chi^2_q.$$

In this expression, $\Sigma_{\varphi\varphi}$ denotes the $q \times q$ submatrix of $\Sigma = J^{-1}KJ$. The matrices J and K were mentioned before. The matrix Σ can be estimated by using the pseudo-likelihood estimate $\widehat{\mathbf{\Phi}}$. Thus, the Wald statistic is very easy to obtain and the more convienient one in cases where model fitting is very time consuming.
The pseudo-score Statistics is constructed by fitting the null model and it has the advantage over the Wald test that is invariant to reparameterisation.
We give another proposal for testing H_0 based on likelihood ratio ideas:

$$G^{*2} = 2[pl(\widehat{\mathbf{\Phi}}) - pl(\varphi_0, \widehat{\beta}(\varphi_0))]$$

and is termed pseudo-likelihood ratio test statistic. Note that, when applying the pseudo-likelihood ratio test, the model needs to be fitted twice, for the full and the reduced models, potentially making the procedure more time consuming. It is well known from the pseudo-likelihood theory that the Wald test is the one with lowest power. However, from a practical point of view it is the more convenient one. All test statistics have been implemented using the SAS IML procedure.

4 Analysis of the Data and Concluding Remarks

The methodology we propose is used to analyse the Moerzeke data, making it the first application of this particular model to data of a genetic type. To proceed we first fit a multivariate Plackett-Dale model and then inferences are made by using the tests proposed.
We restrict the analysis to a subgroup of families having at least one child. From earlier studies we know that for this group, familial transmittance of longevity to daughters is relatively large (Matthijs et al, 2002). In this study, we address whether this association is mainly maternally or paternally transmitted.
We fit the model using the year-of-birth of each family member and the gender of the child as covariates.
The estimated association parameter between mother and child is 1.349 indicating a positive association between those. However, for father-child the value seems to be lower (0.983).
Inferences are made by using the tests defined above. The null hypothesis of no association was tested in each case via the Wald, score and pseudo-likelihood tests and the results show similar conclusions irrespective of the test applied, while the Wald statistics gives the largest p-value. We observe that the null hypotheses of no association between father's and mother's longevities on the one hand and father's and child's longevities on the other hand ($\theta_{12} = 1$ and $\theta_{23} = 1$) but the situation is different for mother and child. Indeed, we reject $\theta_{13} = 1$.
We explore this topic by applying the model to different subsets. We also performed the test for the association parameters for each gender and we can see for sons there are not any significant differences, while for daughters there seems to be a stronger association in case of mothers and daughters than for the rest of the association parameters (θ_{13}).
We proposed three different alternatives to perform inferences for the model parameters: Wald, pseudo-likelihood ratio and score type tests. We illustrated how these test can be performed. Even if the Wald test is the one with less power, in this context we noticed that it is easily implemented from a computational point of view. Even though the pseudo-likelihood and pseudo-score tests are the most powerful, as was observed with other types of data (Geys et al, 1999), here it demands to fit different models. Given the complexity of these models it can be hard to obtain the building blocks needed to calculate these statistics.
The main substantial conclusion is that significant associations were detected between mother and child. In a second step the associations were modelled within the group of daughters and sons separately and we observed significant associations between mother and daughter, but not between mother and sons.
This finding confirms the role of the mother in the transmission of longevity. However, as these findings were present for both mothers and daughters

above the age of fifty as for mothers and daughters reaching at least the age of 10, this finding does not support the view that familial associations in adult mortality are only visible at later ages.

Acknowledgments: The first author wish to thank "Bijzonder Onderzoeksfonds" of the Limburgs Universitair Centrum. The second author thanks to FWO-Vlaanderen. We acknowledge support from Interuniversity Attraction Poles Programme P5/24-Belgian State-Federal Office for Scientific, Technical and Cultural Affairs. We all thank Mr. R. Bijl, a Flemish genealogist and member of the V.V.F. - Dendermonde ("Vlaamse Vereniging voor Familiekunde", Dendermonde division), for allowing us access to his demographic data on Moerzeke.

References

Aerts, M., Geys, H., Molenberghs, G., and Ryan, L. (2002). *Topics in Modelling of Clustered Data*. London: Chapman & Hall.

Arnold, B.C. and Strauss, D. (1991). Pseudo likelihood estimation: some examples. *Sankhya, Series B,* **53**, 233–243.

Dale, J. R. (1986). Global cross–ratio models for bivariate, discrete, ordered responses. *Biometrics,* **42**, 909–917.

Geys, H., Molenberghs, G., and Ryan, L. (1999). Pseudo-likelihood modelling of multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association*, **94**, 34–745.

Le Cessie, S. and Van Houwelingen, J.C. (1994). Logistic regression for correlated binary data. *Applied Statistics*, **43**, 95–108.

Mardia, K.V. (1970). *Families of Bivariate Distributions.* London: Griffin.

Matthijs, K., Van de Putte, B., and Vlietinck, R. (2002). The Inheritance of Longevity in a Flemish Village *European Journal of Population*, **18**, 59–81.

Plackett, R. L. (1965). A class of bivariate distributions. *Journal of the American Statistical Association*, **60**, 516–522.

Renard, D., Molenberghs, G., and Geys, H. (2002). A pairwise likelihood approach to estimation in multilevel probit models, *Computational Statistics and Data Analysis*, **00**, 000-000

Confounding Factors in Time-series Studies of Air Pollution and Health: Effects of Different Adjustment Methods

Touloumi G.[1], Samoli E.[1], Pipikou M.[1], Le Tertre A.[2], and Katsouyanni K.[1] (APHEA-2 project group)

[1] Dept. of Hygiene and Epidemiology, Athens Medical School, Greece
[2] Institute de Veille Sanitaire, Paris, France

Abstract: In time series analysis of air pollution effects on health, several parametric and non-parametric methods to adjust for confounding factors such as trend, seasonality and weather have been proposed. However, an optimal strategy for choosing smoothers and their degree of smoothness does not yet exist. In this paper we evaluate the performance of various smoothers (parametric and non-parametric) with different criteria to choose the degree of smoothness in terms of bias and efficiency in a simulation study. Results showed that non-parametric methods can lead to seriously biased air-pollution effect estimates. Among the parametric methods, Penalized Splines with relatively large number of knots gave minimally biased results. Given that P-splines avoid the backfitting algorithm involved in GAM while being as flexible as the non-parametric smoothing methods it may be a reasonable choice.

Keywords: Air-pollution; GAM; P-splines; Time-series

1 Introduction

Epidemiological time series conducted in cities around the world have reported significant, adverse health effects of air pollution, even at historically low levels of air pollution (Katsouyanni et al 2001; Samet et al 2000). Critics of these studies have raised questions as to the validity of the data, methods of analysis and the rational for particular choices in model specification (Kinney et al 1995; Samet et al 1995). In recent years Generalized Additive Models (GAM; Hastie and Tibshirani, 1990) with non-parametric adjustment for confounding factors have been used to estimate the short-term effects of air pollution on health (Schwartz 1996; Katsouyanni et al 2001). Estimation in GAM is based on a combination of the local scoring algorithm and the backfitting algorithm (Hastie and Tibshirani, 1990) and therefore, unlike Generalized Linear Models (GLM), which have an exact solution, requires iterative approximations. Recently it has been reported that default convergence criteria in statistical packages such as S-plus can

result in biased fitted linear parameters, but this can at least partly overcome by using more stringent convergence criteria (Dominici et al 2002). In another study though it was found that in the presence of concurvity, the nonparametric analogue of multicollinearity, GAMs result in seriously underestimated variances of the fitted model parameters (Ramsay et al 2002). Due to the above-mentioned problems in fitting GAM models, GLM with parametric smoothers (*i.e.* natural splines) for time and weather variables have been alternatively proposed. However, an optimal strategy for choosing smoothers and their degree of smoothness does not yet exist. The aim of this work is to evaluate the performance of various smoothers (parametric and non-parametric) with different criteria for choosing the degree of smoothness in terms of bias and efficiency in a simulation study that imitates multi-center studies.

2 Methods

A two stage hierarchical modeling approach was adopted. In the first stage data from each city were analyzed separately while in the second stage evidence across cities was combined using meta-regression techniques. For the first stage of the analysis Poisson regression models were fitted. The general form of the model was:

$$\ln(\mu_t^c) \;=\; \ln\left[\mathrm{E}\left(Y_t^c\right)\right] \;=\alpha_0 \;+\; \sum_{j=1}^{q} f_j^c(\mathrm{X}_{tj}^c) \;+\beta^c \mathrm{P}_\mathrm{t}^\mathrm{c} + \sum_{i=1}^{6} \beta_i \mathrm{Dow}_i$$

where Y_t^c denotes the observed count of the relevant health outcome (mortality in our case) at city c on day t, β^c the effect estimate for the pollutant (PM_{10} in this case), X_{tj}^c the non-pollution predictor variables (*i.e.*, time, mean daily temperature and mean daily relative humidity), f_j^c smooth functions of these variables and Dow_i indicator variables for the day of the week and μ_t^c is the expected count of the relevant health outcome. To control for non-linear relationships we used the methods: 1) locally weighted non-parametric smoothing (LOESS); 2) smoothing splines (SP); 3) natural splines (NS); 4) penalized - splines (PS; Marx and Eilers 1998; Aerts et al 2002). Each smoother has a parameter that determines the degree of smoothness. For LOESS this is called span while for SP and NS the degree of smoothing is specified through the degrees of freedom. The first two methods are non-parametric smoothers and therefore the backfitting algorithm is needed. The backfitting algorithm cycles through the variables X_{tj} and estimates f_j by smoothing the partial residuals. Except for PS models all the others were fitted in S-plus using either the gam of the glm function for non-parametric and parametric smooth functions respectively. PS models were fitted in R. For the PS models the methodology described by Marx and Eilers (1998) was applied. For the X_{tj} variable the B-Spline

smoother can be specified as $f_j = B_j\alpha_j$, where B_j is the B-Spline matrix (with n_j knots) and α_j the unknown vector of coefficients associated with the B-Spline bases. The P-Spline can be considered as a sum of κ B-Splines, that is $f_j = \sum_{k=1}^{\kappa} B_{jk}\alpha_{kj}$ where $b_{jkt} = B_{jk}(X_{jt})$ is the value of the B-Spline κ at X_{tj}. Marx and Eilers (1998) proposed a smoothness requirement of the B-Spline parameters α_{kj}. A drawback of the method is that one also needs to optimize the number and position of knots. Marx and Eilers (1998) recommended to use a large number of equally spaced knots (between 10 and 30) but prevent over fitting by attaching a difference penalty on adjacent B-Spline coefficients α_{kj}. Model parameters are estimated by maximizing the penalized log likelihood. For LOESS, the span was specified according to the following criteria: a) Minimization of the absolute value of the Sum of the partial autocorrelation function (PACF) over 60 days; b) Minimization of Akaike's information criteria (AIC); c) Minimization of Bayesian information criteria (BIC). For SP the df were prefixed (7/year for time, 6 for temperature and 3 for humidity). The choice of the number of df was based on the results from the NMMAPS study (a multi-city study in the USA; Samet et al 2000). For the NS method, all the three criteria used in the LOESS method plus prespecified df as in the SP method were used. For the PS method the smoothing parameters were determined by a) the generalized cross-validation (GCV) method; b) the GCV after prespecifying the number of knots to be between 10 to 30 for trend and c) equating the needed df to those used in the NMMAPS project.
In the second stage city-specific air pollution effect estimates produced from the first stage of the analysis ($\hat{\beta}^c$) were pooled using inverse variance weighting. Both fixed and random overall estimates were obtained (Berkey et al 1995).

3 Simulation Study

To assess the effect of the various methods to control for confounding effects on city-specific relative rate estimates in time-series studies of air pollution and health we conducted a simulation study. Data were generated under the following assumptions: a) Fifteen cities were contributing daily data for 5 years on daily number of deaths (Y_t^c), PM_{10}^c, mean daily temperature (T^c) and relative humidity (H^c). b) Three different patterns regarding long-term and seasonal trends in mortality series as well as weather conditions were considered. For each pattern 5 cities contributed data. A parametric model was used to generate data in each city (c)

$$Y_t \sim \text{Poisson}(\mu_t)$$

$$\log(\mu_t) = a_0 + (b + b^c)\, PM_{10}^c + \sum_{i=1}^{q} (a_{1i} \cos \tfrac{2\pi i}{365} t + a_{2i} \sin \tfrac{2\pi i}{365} t) + \quad (1)$$
$$+ a_3(\mathrm{T} - T_c)_+^2 + a_4\,(\mathrm{T} - T_c)_+^2 + a_5\,\mathrm{H} + a_6\,\mathrm{H}^2 + \sum_{j=1}^{6} b_j\, Dow_j$$

That is, a sinusoidal curve of order 2,3 and 4 (i.e. q =2 or q = 3, or q = 4) for the three different patterns was used to control for long-term trends, a double quadratic function for temperature (with different change points for the three different patterns) and a quadratic function for humidity. c) Model parameters were generated from the multivariate normal distribution:

$$\begin{pmatrix} b \\ a_i \\ b_j \end{pmatrix} \sim N \left[\begin{pmatrix} \beta \\ \alpha_i \\ \beta_j \end{pmatrix}, V \right]$$

Note that β (the overall PM_{10} effect) is constant and equal to 0.000617 (Katsouyanni et al 2001). An error (b^c) has been added to β in each city to reflect the between cities variability (τ^2), that is $b^c \sim (N(0, \tau^2)$. d) Model parameters were based on real data from the APHEA-2 (A European collaborative study; Katsouyanni et al 2001). In particular, for the three different patterns data from 3 cities (London, Cracow and Madrid) representative of the different geographical areas across Europe (North, Central and South Europe, respectively) were used. True values of the model parameters were determined by fitting model (1) in these 3 cities. The original data on temperature humidity and PM_{10} from these cities (to incorporate original structures in the data) were used as the covariates in the simulated data.

4 Results

The PM_{10} pooled estimate was sensitive to the method used to adjust for confounding factors such as trend, seasonality and weather. Depending on the method, the bias ranged from -27% to 17%. Within each method, the different criteria to choose the appropriate df led to different degree of bias in the pooled PM_{10} estimate. In addition, non-parametric methods (*i.e.*, LOESS and SP) tended to underestimate the SE of the PM_{10} effect. For the P-spline method, using GCV to determine the smoothness parameters led to underfitting (*i.e.*, relative small dfs) and therefore to seriously biased pooled PM_{10} estimate (bias -16%). However, improving model fit by increasing the dfs (*i.e.*, prespecifying either the total dfs or the number of knots) resulted in substantial bias reduction (bias -3% and -7% respectively).

5 Conclusions

In time series analysis of air pollution effects on health, some degree of concurvity is expected. In such cases, GAM models could lead to biased

results. Among the parametric methods to adjust for confounding factors, P-spline (with relatively large number of knots, for example around 30) avoids the backfitting algorithm involved in GAM, and thus all the associated weaknesses, while being as flexible as the non-parametric smoothing methods.

References

Aerts, M., Claeskens, G., and Wand, M.P. (2002). Some theory for penalized spline generalized additive models. *Journal of Statistical planning and inference*, **103**(1-2), 455–470.

Berkey, C.S., Hoaglin, D.C., Mosteller, F., and Colditz, G.A. (1995). A random-effects regression model for meta-analysis. *Statistics in Medicine*, **14**, 395–411.

Dominici, F., McDermott, A., Zeger, S.L., and Samet, J.M. (2002). A cautionary note regarding generalized additive models software: Implications for time-series studies of air pollution and health. *American Journal of Epidemiology*. (in press).

Hastie, T.J., and Tibshirani, R.J. (1990). *Generalized additive models*. New York: Chapman & Hall.

Katsouyanni, K., Touloumi, G., Samoli, E., et al (2001). Confounding and effect modification in the short-term effects of ambient particles on total mortality: Results from 29 European cities within the APHEA2 project. *Epidemiology*, **12**, 521–531.

Kinney, P.L., Ito, K., and Thurston, G.D. (1995). A sensitivity analysis of mortality /PM_{10} associations in Los Angeles. *Inhalation Toxicology*, **7**, 59–69.

Marx, B.D. and Eilers, P. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics and Data Analysis*, **28**, 193–209.

Ramsay, T., Burnett, R., and Krewski, D. (2002). The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. *Epidemiology*, **14**, 18–23.

Samet, J.M., Dominici, F., Curricro, F.C., Coursac, I., and Zeger, S.L. (2000). Fine particulate air pollution and mortality in 20 U.S. cities,1987–1994. *New England Journal of Medicine*, **343**, 1742–1749.

Samet, J.M., Zeger, S.l., and Berhane, K. (1995). The association of mortality and particulate air pollution. *Particulate Air Pollution and Daily Mortality: Replication and Validation of Selected Studies*. 1–104, Health Effects Institute, Cambridge MA.

Schwartz, J., Dockery, D.W., and Neas, L.M. (1996a). Is daily mortality associated specifically with fine particles? *Journal of the Air & Waste Management Association*, **46**, 2–14.

Bayesian Parsimonious Estimation of Observed and Unobserved Heterogeneity

Regina Tüchler[1] and Sylvia Frühwirth-Schnatter[2]

[1] Department of Statistics, University of Economics and Business Administration, Augasse 2-6, A-1090 Vienna, Austria. Email: regina.tuechler@wu-wien.ac.at

[2] Department of Applied Statistics (IFAS), Johannes Kepler Universität, Altenbergerstr. 69, A-4040 Linz, Austria, Email: Sylvia.Fruehwirth-Schnatter@jku.at

Abstract: In this work we use MCMC methods to estimate random coefficient models. The following two issues which we believe are of considerable practical concern are addressed. Firstly, the convergence behaviour depends on the parameterization. An inappropriate parameterization may have a serious impact on the mixing properties especially for higher dimensional data. Secondly, a mayor cause for poor convergence of MCMC chains stems from the attempt to estimate over-fitted models. We present an efficient algorithm that makes it possible to start with a rather general model structure and to let the data tell us which special structure to choose.

Keywords: Covariance matrix; Variable selection; Random coefficient model, Heterogeneity Model; MCMC Methods.

1 The Random Coefficient Model

In the *centered* parameterization we define the random coefficient model in the following way:

$$y_i = Z_i\beta_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2 I), \tag{1}$$

$$\beta_i = \beta^G + u_i, \quad u_i \sim N(0, Q). \tag{2}$$

For subjects $i = 1, \ldots, N$ the vector y_i contains T_i observations and Z_i is the $T_i \times d$-dimensional design matrix for the d individual effects β_i. The vectors u_i capture *unobserved heterogeneity* as deviations of the individual effects β_i from the common mean β^G and are distributed normally with common covariance matrix Q. The model (1), (2) is *centered* both in the mean and in the covariance structure.
We may rewrite that model in the *non-centered* parameterization as:

$$y_i = Z_i\beta^G + Z_iC\tilde{z}_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2 I), \tag{3}$$

$$\tilde{z}_i \sim N(0, I), \tag{4}$$

where the Cholesky decomposition of the covariance matrix $Q = CC'$ is used, C lower triangular. The Cholesky decomposition is a representation of the correlated deviations u_{il} from (2) in terms of uncorrelated standard normal ones, $u_{il} = \sum_{m=1}^{l} C_{lm}\tilde{z}_{im} = C_{l\cdot}\tilde{z}_i$, where $\tilde{z}_{im} \sim N(0,1)$ are independent for all $m = 1, \ldots, d$, and $\dim(u_i) = \dim(\tilde{z}_i)$. In the *non-centered* parameterization the mean β^G and the covariance Q moved to the observation equation (3).
The influence of the parameterization of the mean on the convergence behaviour was analyzed by Gelfand, Sahu and Carlin (1995) for normal linear models.
Non-centering of the mean and the covariance is investigated in Meng and Van Dyck (1998) and Frühwirth-Schnatter (2002).

2 Parsimonious Estimation of Unobserved Heterogeneity

Statistical inference for the covariances of a random coefficient model is usually based on the estimation of a full rank matrix Q from (2) (see e.g. Verbeke and Lesaffre, 1996, Frühwirth-Schnatter, Tüchler and Otter, 2003). Within MCMC sampling the covariance matrix Q may be simulated by a Gibbs sampler from the conditional inverted Wishart distribution given estimates of $\beta_1, \ldots, \beta_N$. Note that by choosing a full rank matrix Q, we allow unobserved heterogeneity to be present for all effects of the design Z_i.
In contrast to that, we may follow the principle of adaptive parsimony with respect to Q. Parsimony is achieved by restricting certain elements appearing in the matrix of the Cholesky factors C to be 0. We let the data tell us which elements this should be. We treat the problem of finding those elements of C that are non-zero as a variable selection problem. We introduce for each element C_{lm}, $m = 1, \ldots, d$, $l = m, \ldots, d$, an indicator γ_{lm} which takes the value 1, if $C_{lm} \neq 0$, and 0 otherwise:

$$\begin{aligned} C_{lm} &= 0, \quad \text{iff } \gamma_{lm} = 0, \\ C_{lm} &\neq 0, \quad \text{iff } \gamma_{lm} = 1. \end{aligned}$$

Note that C_{lm} for all $1 \leq l < m$ is 0 by definition. $\gamma = \{\gamma_{lm}\}$, as well as the omitted variables $\tilde{z}^N = (\tilde{z}_1, \ldots, \tilde{z}_N)$ are estimated along with all other parameters in a Bayesian framework using MCMC methods.
The problem of finding the form of the covariance matrix is closely related to variable selection problems and therefore to the issue of estimating observed heterogeneity. For these problems we can think of C as parameter matrix and $\tilde{z}_i$ would be deterministic in equation (4). We think that it is of considerable practical concern that variable selection problems may easily be treated within a Bayesian framework by obvious extensions of the methods of this paper.

3 Bayesian Estimation According to the Principle of Adaptive Parsimony

We estimate all parameters from their joint posterior distribution using a Markov chain Monte Carlo algorithm. In the first step (I) we generate the indicators γ_{lm} one at a time from $\gamma_{lm}|\gamma_{\backslash lm}, \tilde{z}^N, \sigma_\varepsilon^2, y$ by applying the efficient sampling scheme of Smith and Kohn (2002). $\gamma_{\backslash lm}$ denotes the sequence γ where γ_{lm} is excluded and y are the data. Once a new draw of γ is available, all those elements of C that are restricted to zero are defined. In step (II) we generate all unrestricted elements of C jointly with β^G conditionally on $\gamma, \tilde{z}_i, \sigma_\varepsilon^2$ and y from a multivariate normal distribution. In step (III) the individual parameters $\tilde{z}_1, \ldots, \tilde{z}_N$ are conditionally independent and normally distributed. Step (IV) amounts to sampling σ_ε^2 from inverted gamma distributions.

4 Example

We simulated data for $N = 200$ subjects from model (1), (2) with $\beta^G = (15\ 3\ -0.8)'$, $\sigma_\varepsilon^2 = 10$ and a sparse covariance matrix

$$Q = \begin{pmatrix} 2.25 & 0 & 3 \\ 0 & 0 & 0 \\ 3 & 0 & 5.25 \end{pmatrix}. \tag{5}$$

We estimated these data with three different algorithms. To compare the convergence properties of the various methods we give the sample paths and the autocorrelation plots for the first component of the mean and the first diagonal element of the covariance matrix, denoted beta1 and q1 in the plots, respectively.

Full conditional Gibbs sampling, centered parameterization: The first method is a full conditional Gibbs sampling algorithm based on the centered parameterization (1), (2). The covariance matrix is estimated from inverted Wishart. In Figure 1 we see that the algorithm is rather inefficient in terms of autocorrelations.

Full conditional Gibbs sampling, non-centered parameterization, model (3), (4)*, no variable selection:* In comparison to the sample paths and autocorrelation plots of Figure 1 those for the non-centered parameterization in Figure 2 are much better at least for the mean structure. This is inline with the results in Gelfand et al (1995) where it is stated that the non-centered parameterization is to be preferred for data with little heterogeneity in the random effects in comparison to the variability captured by the model error.

Parsimonious estimation of the covariances: As we can see in Figure 3 parsimonious estimation of the covariance matrix improves the convergence behaviour of the sampler substantially. The algorithm finds the correct

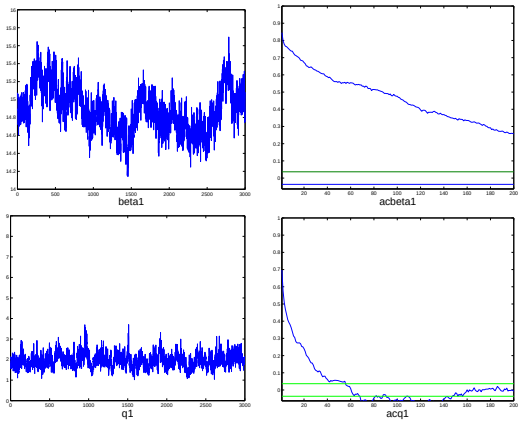


FIGURE 1. *Full conditional Gibbs sampling, centered parameterization*

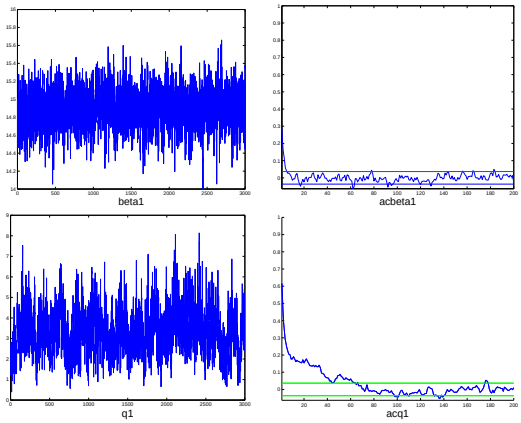


FIGURE 2. *Full conditional Gibbs sampling, non-centered parameterization*

structure of the covariances very well. The following mean indicator may be interpreted as probability of an element of Q to be significant:

$$prob(Q \neq 0) \quad = \quad \begin{pmatrix} 1 & 0.00 & 1 \\ 0.00 & 0.01 & 0.00 \\ 1 & 0.00 & 1 \end{pmatrix}.$$

The second effect is no random effect but a fixed effect. This feature may not be detected by the first two algorithms which sample a wrong full rank covariance matrix Q. This overparameterization has also a negative effect on the convergence behaviour of the other (significant) elements of Q.

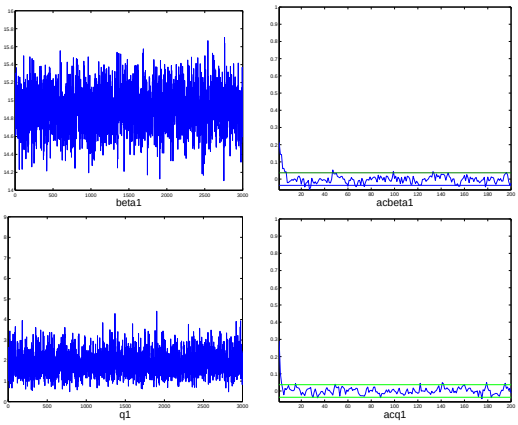


FIGURE 3. *Parsimonious estimation of the covariance matrix, non-centered parameterization*

Acknowledgments: This work was supported by the Austrian Science Foundation (FWF) under grant SFB 010 ('Adaptive Information Systems and Modeling in Economics and Management Science').

References

Frühwirth-Schnatter, S. (2002). Efficient Bayesian parameter estimation for state space models based on reparameterizations. Working paper, Department of Statistics, Vienna University of Economics and Business Administration, No. 2002-83.

Frühwirth-Schnatter, S., Tüchler, R., and Otter, T. (2003). Bayesian analysis of the heterogeneity model. *Journal of Business and Economic Statistics*. (in press).

Gelfand, A.E., Sahu, S.K., and Carlin, B.P. (1995). Efficient parameterizations for normal linear mixed models. *Biometrika*, **82**, 479–488.

Meng, X.-L. and van Dyk, D. (1998). Fast EM-type implementations for mixed effects models. *Journal of Royal Statistical Society, Series B*, **60**, 559–578.

Smith, M. and Kohn, R. (2002). Bayesian parsimonious covariance matrix estimation. *Journal of the American Statistical Association*. (in press).

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, **91**, 217–221.

Use of a Mixed model to Estimate the Number of People who Died Because they had HIV/AIDS

Joanne Tyler

[1] Department of Statistics, University of Fort Hare, Alice 5700, South Africa. Email: jtyler@ufh.ac.za

Abstract: It is very difficult in Third World countries to obtain reliable statistics about the numbers of people who die because they have HIV/AIDS. This paper attempts to use conditional distributions to find out the numbers of people who die from the opportunistic diseases associated with HIV/AIDS and for whom it should have been noted that HIV/AIDS was a pre-existing condition.

Keywords: Conditional survival analysis; HIV/AIDS; Competing risks.

1 Introduction

In many countries, particularly in the Third World, it is very difficult to obtain reliable data on HIV/AIDS because most death certificates note only the opportunistic disease e.g.TB, meningitis, pneumonia which actually caused the patient to die, while making no or infrequent mention of the pre-existing condition of HIV/AIDS. One frequently sees competing risks models based on the proportional hazards model. Here it will be shown that, since HIV/AIDS data fits an accelerated failure time model better than a proportional hazards model, a parametric survival function will be used with logistic regression to perform the function of a mixed model in modeling HIV/AIDS data.

2 Conditional Accelerated Failure Time Models as Functions of Cause Specific Hazard Rate Functions, $\boldsymbol{\alpha_i(t)(i = 1, 2)}$.

Suppose we consider a 2-cause competing risks model with cause specific hazard rate functions $\alpha_i(t)$ $(i = 1, 2)$ where $\alpha_i(t) = p_i(t)/S(t)$, $p_i(t)$ being the pdf of $\alpha(t)$ and $S(t)$ being the survivor function. According to Hougaard (2000), a competing risks model allows only one possible transition into each state. This would mean that a competing risks model is not a multi-state model. This introduces an interesting concept in the case

of HIV/AIDS. Before any treatment has occurred we have a multi-state model and therefore a competing risks model but once treatment by anti-retroviral drugs has taken place we can no longer have a competing risks model as can be seen from the chart below. This chart imitates a poster at IBS by Heiner (2002).
The overall hazard rate function will then be

$$\alpha(t) \;=\; \alpha_1(t) + \alpha_2(t)$$

and given that death has already occurred, the conditional probability that it was due to cause (i) is given by

$$\pi_i \;=\; \frac{\alpha_i(t)}{\alpha(t)}$$

In modelling the hazard rate associated with cause (i) ($i = 1, 2$), each person is treated as censored at the occurrence of death from any cause other than cause (i). Accelerated failure time models can be formulated in terms of the hazard rate function (Bagdonavicius and Nikulin, 2002). It is generally accepted that how the overall hazard rate function divides itself amongst its constituent parts can be seen by analyzing the conditional probabilities of each of the hazard rates but is in itself not of prime importance. What is of greatest importance is the cause of death. We can therefore write

$$S(t) \;=\; \exp\left\{-\int_0^t \alpha(u)du\right\}$$

To analyze the conditional probabilities of each of the hazard rate a hazard model for the overall hazard rate is specified and analyzed along with a logistic regression of the conditional probability of the cause of death of interest (Ghilagaber, 1998).
It can be shown that from this it follows that

$$S(t,z) \;=\; \exp\left\{-\int_0^t \alpha(u)du\right\}\exp\{\beta z\}$$

and so for each cause specific hazard function we have

$$S(t/\alpha_i(t)) \;=\; \exp\left(-\int_0^t \alpha_i(u)du\right) \qquad (i = 1, 2)$$

and the probability $P(t/\alpha_i(t,z) \;=\; p_i(t,z)$
where a logistic regression model for $\pi_1(t,z)$ is:

$$logit[\pi_1(t,z)] \;=\; \ell n\left\{\frac{\pi_1(t,z)}{1-\pi_1(t,z)}\right\} \;=\; \beta_0(t) + \beta z \qquad (1)$$

There are thus proportional hazards over death from opportunistic diseases but not over the covariates.

TABLE 1. *Number of male deaths*

HIV/AIDS	TB	Pneu-monia	Karposi's Sarcoma	Gastric diseases	Meningitis	Other
17	2	0	0	0	0	0
13	9	3	0	0	1	0
23	0	0	3	14	20	66

TABLE 2. *Number of female deaths*

HIV/AIDS	TB	Pneu-monia	Karposi's Sarcoma	Gastric diseases	Meningitis	Other
40	6	0	0	4	1	1
51	8	4	0	0	2	0
76	0	0	1	10	22	1

3 Data Analysis

Data collected at three local hospitals yielded the following causes of death. Patients had been tested for HIV/AIDS and the other diseases can therefore be classified as opportunistic diseases.
We see from the tables that while 220 people were registered as having died of HIV/AIDS during a given period of time, if the numbers of people who were known to have HIV/AIDS as well as the opportunistic diseases from which they died then the actual number of people who died as a result of having HIV/AIDS during the given period of time was in fact 398.

4 Summary

We have shown that death from opportunistic diseases known to be associated with HIV/AIDS is closely related to the pre-existing condition of HIV/AIDS. See equation (1). This means that the number of deaths from these diseases should be taken into account when compiling the statistics of death from HIV/AIDS. It is hoped that it would then be possible to obtain a more realistic idea of how many people have died as a result of HIV/AIDS, particularly in the Third World countries. It can be shown that the Weibull is the only initial distribution for which with constant explanatory variables the accelerated life models and the proportional hazards models coincide (Cox and Oakes, 1984).

References

Bagdonavicius, V. and Nikulin, M. (2002). *Accelerated Life Models*. Chapman & Hall.

Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall/CRC.

Ghilagaber, G. (1998). Analysis of survival data with multiple causes of failure: a comparison of hazard- and logistic-regression models with application in demography. *Quality and Quantity*, **32**, 297–324.

Heiner, K. (2002). Describing response to HAART in a public health care setting. IBS: Poster session.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Heidelberg: Springer-Verlag.

Evaluation of PQL in Disease Mapping

M.D. Ugarte[1], A.F. Militino[1], and C.B. Dean[2]

[1] Departamento de Estadística e Investigación Operativa, Campus de Arrosadía, Universidad Pública de Navarra, 31006 Pamplona, Navarra, Spain
[2] Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, B.C. V5A 1S6 Canada

Abstract: This work discusses and evaluates penalized quasi-likelihood estimation techniques for the situation where random effects are correlated, as is typical in mapping studies.

Keywords: Mixed poisson model; Over-dispersion; Spatial autocorrelation.

1 Introduction

Full maximum likelihood analysis in generalized linear mixed models usually involves iterative numerical quadrature. Breslow and Clayton (1993), however, have popularized the use of penalized quasi-likelihood (PQL) methods developed by Stiratelli et al (1984) and Schall (1991) for inference in these models. PQL analysis relies on a series of approximations to the mixed model. Its main advantage is that its implementation is very straightforward and computationally simple. Estimation proceeds using a so-called working vector and the restricted maximum likelihood (REML) equations under the normal theory linear model. Lin (1994) has shown, through a study of the theoretical properties of PQL estimators, that they are reliable for the analysis of independent counts. For the analysis of proportions with low denominators, Breslow and Lin (1995) note that PQL estimators show substantial bias.

In this paper we aim to assess the suitability of PQL inference for random effects models where the random effects are correlated. The setting we use for our illustration and study is that of mapping studies. Mapping rates is essentially a problem of describing the spatial and sometimes spatio-temporal distribution of rates over a region. Such distributions display the geographic variation in mortality or disease incidence and are important for epidemiological and health-policy purposes. The models in common use for assessing spatial variation are mixed Poisson models, which incorporate random region-specific effects. These region-effects may represent environmental factors, and often exhibit spatial correlation.

In this work a correlated mixed Poisson model and penalized quasi-likelihood estimation for this model will be described. The model includes

an interaction between random regions and fixed age effects (Dean, Ugarte and Militino, 2001). An analysis of mortality data from British Columbia (B.C.), Canada, over the five-year period 1985-1989 will be presented. We will also evaluate the behaviour of PQL estimators for small samples using the basic scenario as encountered in the analysis of the B.C. mortality data, but also encompassing different levels of spatial correlation and reduced population sizes. Our results show that the PQL estimators have fair small-sample properties. When mortality counts are small, however, our simulations show a greater deterioration in the performance of the PQL estimators of the variance components than that for the parameters in the mean.

2 Description of the Model

Suppose the area under study is divided into I contiguous regions labelled $i = 1, \ldots, I$. In British Columbia, these are termed *local health areas*, and $I = 79$. Let C_{ij} be the number of stratum-specific (e.g. sex, disease) deaths in the i-th region for the j-th age group $j = 1, \ldots, J$. Conditional on the random region effects, the number of deaths in each area is assumed to be Poisson distributed with mean $\mu_{ij} = e_{ij} r_{ij}$, where r_{ij} are the unknown relative risks of mortality from the disease, treated as random effects, and e_{ij} are the 'expected' number of deaths. That is, given the random effects r_{ij},

$$C_{ij} | r_{ij} \sim \text{Poisson}(e_{ij} r_{ij}) \tag{1}$$

where $e_{ij} = n_{ij} m_j$, n_{ij} being the corresponding population count in the time period considered and m_j being a fixed age effect representing the mean mortality rate for the j-th age-group. We decompose the region by age-group log-relative risks as the sum of two independent components:

$$\log r_{ij} = u_i + w_{ij}. \tag{2}$$

The term u_i models both intrinsic Gaussian autoregression (Besag, 1974), representing local spatially structured variation, and the unstructured heterogeneity of the relative risks which is usually associated with covariates not included in the model. Here $\mathbf{u} \sim N(\mathbf{0}, \mathbf{D}_u), \mathbf{D}_u = \sigma_u^2(\lambda \mathbf{Q}^- + (1-\lambda)\mathbf{I_u})$, where $\mathbf{Q}$ is determined by the neighborhood structure of the regions and $\mathbf{I_u}$ is an identity matrix; λ determines the relative weights between the spatial and the unstructured variation. When $\lambda = 1$, there is no unstructured heterogeneity, and the random effect u_i can be interpreted conditionally given u_{-i}, the set of random region-effects excluding the ith:

$$u_i | u_{-i} \sim N(\bar{u}_{\delta_i}; \sigma_u^2/\delta_i), \tag{3}$$

$\bar{u}_{\delta_i}$ is the mean of the random effects corresponding to the regions in the 'neighborhood' of the ith and δ_i is the number of regions forming this neighborhood. Neighborhoods may be defined in various ways, depending on the context of the analysis, but one common definition is simply the set of regions which share common boundaries with the ith region, and this definition is adopted in our analyses described later. In this case $\mathbf{Q}$ has ith diagonal element equal to the number of neighbors of the ith region and the off-diagonal elements of each row equal -1 if the corresponding regions are neighbors and 0 otherwise. The more general formulation for user-specified weights ψ_{ij} linking regions i and j is

$$u_i|u_{-i} \sim N\left(\frac{\sum_j \psi_{ij} u_j}{\sum_j \psi_{ij}}, \frac{\sigma_u^2}{\sum_j \psi_{ij}}\right),$$

that is, the conditional mean of u_i is a weighted mean of the other region-effects. The conditional distribution (3) is appropriate for the 0/1 adjacency weights used here. When $\lambda = 0$, there is no spatial correlation in the data and $u_i \sim N(0, \sigma_u^2)$ independently. Because the specification is in terms of conditional distributions, $\mathbf{Q}$ is a singular matrix (Besag, 1974; Besag and Kooperberg, 1995). In generalized linear modeling, the typical solution would be to reduce the problem to that of estimating a reduced set of random effects with full rank variance matrix. This can be equivalently handled using the Moore-Penrose generalized inverse, $\mathbf{Q}^-$ (Harville, 1997, chp. 20), which is the approach used in our applications.

The term w_{ij} in (2) represents an age-region interaction term, $w_{ij} \sim N(0, \sigma_w^2)$, independent of u_i, $i = 1, \ldots, I$.

PQL estimation is an approximate inference technique for generalized linear mixed models which uses weighted least squares algorithms for estimation of parameters in the mean along with likelihood equations from an approximating normal model for estimating variance components. In terms of a generalized linear mixed model, (1) and (2) can be expressed more generally as $E(\mathbf{C}|\mathbf{b}) = \mu^b = g^{-1}(\text{offset} + X\alpha + Zb)$, where $\mathbf{C}$ is the vector of responses, in our situation, mortality counts C_{ij}; α is the vector of fixed effects, here $\alpha = (\log m_j, j = 1, \ldots, J)$; X is the corresponding design matrix here of dimension IJ x J; the offset for the mapping scenario is the known vector of the logarithm of the population counts n_{ij}; b is the vector of random effects, here $b = (u^T, w^T)^T$, $Zb = Z_1 u + Z_2 w$, Z_1 and Z_2 being corresponding design matrices having dimensions IJ x I and IJ square respectively, with Z_2 being an identity matrix. The function $g^{-1}(.)$ is the inverse of the so-called link function, so $g(\mu^b)$ is the linear predictor η. For the log-linear specification used here, $\eta = \log(\mu^b)$, so g^{-1} is the exponential function, and $\eta_{ij} = \log \mu_{ij} = \log n_{ij} + \log m_j + u_i + w_{ij}$ (see (2.1) and (2.2)).

The integrated quasi-likelihood function is

$$|D|^{-1/2} \int \exp\left[-\frac{1}{2}\sum_{i,j} d_{ij}(y_{ij}, \mu_{ij}^b) - \frac{1}{2}\, b^T D^{-1} b\right] db$$

with

$$d(y, \mu) = -2 \int_y^\mu \frac{y-u}{u}\, du$$

and $\text{var}\,(b) = D$,

$$D = \begin{pmatrix} D_u & \mathbf{0} \\ \mathbf{0} & \sigma_w^2 I_w \end{pmatrix},$$

where I_w is an identity matrix of dimension IJ. Breslow and Clayton (1993) exploited ideas on Laplace methods for integral approximations. By taking a quadratic expansion of the term in the exponent about its maximizing value before integration, the penalized quasi-likelihood is obtained. This leads to a Laplace approximation to the integrated quasi-likelihood.
To correspond closely with the normal mixed effects model so that an iterative weighted least squares algorithm may be applied to estimate the fixed effects, Breslow and Clayton (1993) define a working response vector Y to have components $Y_{ij} = \eta_{ij} - \text{offset} + (C_{ij} - \mu_{ij}) g'(\mu_{ij})$, where g' is the derivative of g; here $g'(\mu_{ij}) = 1/\mu_{ij}$. The associated normal theory model is then

$$Y = X\alpha + Zb + \epsilon, \tag{4}$$

where $\epsilon \sim N(\mathbf{0}, W^{-1})$, $W = \text{diag}\{\text{var}(C_{ij}|b)[g'(\mu_{ij})]^2\}^{-1}$. The estimate of α is obtained as $\hat{\alpha} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y$, with estimated asymptotic variance $(X^T V^{-1} X)^{-1}$, $V = W^{-1} + ZDZ^T$. In our mapping context, $V = W^{-1} + Z_1 D_u Z_1^T + \sigma_w^2 I_w$ and $W = \text{diag}\,(\mu_{ij})$. Note that this estimated asymptotic variance does not account for the additional variability which results in the typical scenario where variance components are estimated, hence standard errors of the $\hat{m}_j$'s may be underestimated. Random effects are estimated as empirical Bayes estimates of their posterior mean given the data: $\hat{b} = DZ^T V^{-1}(Y - X\hat{\alpha})$.
For estimation of the variance components, the restricted maximum likelihood (REML) equations are employed (Harville, 1977),

$$\frac{1}{2}\left[(Y - X\hat{\alpha})^T V^{-1} \frac{\partial V}{\partial \theta_r} V^{-1} (Y - X\hat{\alpha}) - \text{tr}\left(P \frac{\partial V}{\partial \theta_r}\right)\right] = 0, \qquad r = 1, 2, 3, \tag{5}$$

where $P = V^{-1/2}(I - H)V^{-1/2}$, H being a projection matrix typically called the *hat* matrix, $H = V^{-1/2}X(X^TV^{-1}X)^{-1}X^TV^{-1/2}$. The asymptotic variance of $\hat{\theta}$ is $\mathcal{I}^{-1}$, where $\mathcal{I}$ has components

$$\mathcal{I}_{rs} = (1/2)\text{tr}\,[P(\partial V/\partial\theta_r)P(\partial V/\partial\theta_s)],\ \ r, s = 1, 2, 3.$$

3 Results

Penalized quasi-likelihood is a very straightforward technique to implement for GLMMs. Our simulation results (not shown here. See Dean, Ugarte and Militino, 2003) indicate that PQL estimators have little bias for the correlated model considered, except when means are very low yielding sparse data with an abundance of zero counts. These results suggest that PQL may be reliably used for exploratory studies, such as routine map production at health agencies. Other approaches, e.g., Markov Chain Monte Carlo methods, are available for etiological studies and other detailed analyses.

Acknowledgments: The research of Ugarte and Militino has been supported by the Health Department of the Government of Navarra, Spain (project Res. 1878/2001). Dean's research was supported by the Natural Sciences and Engineering Research Council of Canada.

References

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

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

Breslow, N.E. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, **82**, 81–91.

Dean, C.B., Ugarte, M.D., and Militino, A.F. (2001). Detecting interaction between random regions and fixed age effects in disease mapping. *Biometrics*, **57**, 197–202.

Dean, C.B., Ugarte, M.D., and Militino, A.F. (2003). Penalized quasi likelihood with spatially correlated data. *Computational Statistics and Data Analysis*. (in press).

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**, 320–340.

Lin, X. (1994). *Bias correction in generalized linear mixed models*. Ph.D. thesis, University of Washington, Seattle.

Schall, R. (1991). Estimation in generalized linear models with random effects, *Biometrika*, **78**, 719–727.

Stiratelli, R., Laird, N., and Ware, J. (1984). Random effects models with serial observations with binary responses, *Biometrics*, **40**, 719–727.

Testing Homogeneity in Weibull Error in Variables Models

Dione M. Valença[1] and Heleno Bolfarine[2]

[1] Departamento de Estatística, Universidade Federal do Rio Grande do Norte, Cep 59072-970 - Natal - RN - Brazil, Email: dione@ccet.ufrn.br
[2] Instituto de Matemática e Estatística, USP - Caixa Postal 66281 (Ag. Cidade de São Paulo), 05311-970 São Paulo - SP - Brazil, Email: hbolfar@ime.usp.br

Abstract: We discuss properties of the score statistics for testing the null hypothesis of homogeneity in a Weibull mixing model in which the group effect is modeled as a random variable and some of the covariates are measured with error. The statistics proposed are based on the corrected score approach and they require estimation only under the conventional Weibull model with measurement errors and does not require that the distribution of the random effect be specified. The results in this paper extend results in Gimenez et al. (2000) for the case of independent Weibull models. A simulation study is provided.

Keywords: Homogeneity test; Measurement errors; Corrected score; Accelerated failure time model.

1 Weibull Measurement Error Models with a Random Effect

Consider a sample divided into k groups and let T_{ij} be the event time to the individual j in the group i, with $j = 1, \ldots, n_i$, and $i = 1, \ldots, k$. The log-linear Weibull model with a random effect, models $\log T_{ij}$ as

$$logT_{ij} = U_i + \boldsymbol{\beta}_z^{\mathsf{T}} \mathbf{z}_{ij} + \beta_x x_{ij} + \sigma\epsilon_{ij}, \tag{1}$$

where the $\epsilon'_{ij}s$ are independent and identically distributed ($i.i.d.$) random errors with standard extreme value density function $f(\epsilon) = \exp(\epsilon - e^{\epsilon})$, $\epsilon \in \Re$. We consider $\mathbf{z}_{ij}$ a covariate vector correctly observed and x_{ij} an unobserved variable which is measured with error. We assume an additive functional measurement error model relating the observed (surrogate) w_{ij} and the unobserved x_{ij}, which is expressed as $w_{ij} = x_{ij} + \eta_{ij}$, where $\eta'_{ij}s$ represent unobserved $i.i.d.$ errors with distribution $N(0, \phi)$. The random effect for group i is represented as

$$U_i = \alpha + \theta^{1/2} V_i,$$

where the $V_i's$ are *i.i.d.* random variables with, $E[V_i] = E[V_i^3] = 0$, $E[V_i^2] = 1$, and $E[V_i^m] < \infty$, $m > 3$, and otherwise unspecified distribution function F. We assume that $U_i, \eta_{ij}, \epsilon_{ij}$ are all independent.

2 Marginal Likelihood

Consider that survival times are subject to right censoring and that censoring is random, uninformative and independent of U_i. Set $\delta_{ij} = 1$ to indicate a failure time and $\delta_{ij} = 0$ to indicate a censoring time. Let Y_{ij} be the observed log survival time for subject j in group i. Denote by $\lambda = (\boldsymbol{\gamma}^\mathsf{T}, \theta)^\mathsf{T}$, the vector of parameters, with $\gamma^\mathsf{T} = (\alpha, \boldsymbol{\beta}_z^\mathsf{T}, \beta_x, \sigma)$. The hypothesis of homogeneity is $H_0 : \theta = 0$. The likelihood with respect to the conditional distribution of (Y_{ij}, δ_{ij}) given V_i is

$$L_{ij}(\lambda|u(v_i), x_{ij}) = (1/\sigma)^{\delta_{ij}} \exp[\delta_{ij} s(x_{ij}, v_i) - \exp(s(x_{ij}, v_i))], \qquad (2)$$

where $u(v_i) = \alpha - \theta^{1/2} v_i$ and $s(x_{ij}, v_i) = (y_{ij} - u(v_i) - \boldsymbol{\beta}_z^T \mathbf{z}_{ij} - \beta_x x_{ij})/\sigma$. Note that for $v_i = 0$, $L_{ij}(\lambda|u(0), x_{ij}) = L_{ij}(\gamma, x_{ij})$ depends only of γ. Let $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})^\mathsf{T}$ and $\mathbf{Y} = (Y_1^\mathsf{T}, ..., Y_k^\mathsf{T})$, and assume other vectors denoted similarly. The marginal log-likelihood corresponding to the observed sample is given by

$$l(\lambda, \mathbf{x}) = \sum_{i=1}^{k} log \int \prod_{j=1}^{n_i} L_{ij}(\lambda|u(v_i), x_{ij}) dF(v_i). \qquad (3)$$

The solution proposed in Bolfarine and Valença (2002), follows by making a Taylor series expansion of the integrand in (3) around $v_i = 0$, which leads to

$$l(\lambda, \mathbf{x}) = l_0(\gamma, \mathbf{x}) + \sum_{i=1}^{k} \log \left[1 + \frac{h_i(\gamma, x_i)\theta}{2} + \sum_{m=3}^{\infty} \frac{D_i^{(m)}(\gamma, x_i) E(V_i^m) \theta^{\frac{m}{2}}}{m!} \right],$$

$$D_i^{(m)}(\gamma, x_i) = \left(\frac{\partial^m L_i(\gamma, x_i)}{\partial \alpha^m} \right) / L_i(\gamma, x_i) \qquad (4)$$

where $L_i(\gamma, x_i) = \prod_{j=1}^{n_i} L_{ij}(\gamma, x_{ij})$ and $l_0(\gamma) = \sum_{i=1}^{k} \log L_i(\gamma, x_i)$, with $L_{ij}(\gamma, x_{ij}) = L_{ij}(\lambda|u(0), x_{ij})$ given in (2). Denote $s(x_{ij}) = s(x_{ij}, 0) = (y_{ij} - \alpha - \boldsymbol{\beta}_z^\mathsf{T} \mathbf{z}_{ij} - \beta_x x_{ij})/\sigma$. The quantity $h_i(\gamma, x_i) = D_i^{(2)}(\gamma, x_i)$, can be written as

$$h_i(\gamma, x_i) = \frac{1}{\sigma^2} \left\{ \left[\sum_{j=1}^{n_i} (\exp(s(x_{ij})) - \delta_{ij}) \right]^2 - \sum_{j=1}^{n_i} \exp(s(x_{ij})) \right\}, \qquad (5)$$

3 Naive Tests of Homogeneity

Let $\mathsf{S}(\lambda;\mathbf{x}) = \partial l(\lambda,\mathbf{x})/\partial\lambda$ be the score function and $I(\lambda;\mathbf{x}) = \frac{-\partial^2 l(\lambda,\mathbf{x})}{\partial\lambda\partial\lambda^{\mathsf{T}}}$ be the observed information matrix. Note that S and I can not be computed when the true x_{ij} is measured with error since it is not observed. One alternative is to replace the unobserved x_{ij} by the observed w_{ij}, and ignore the measurement error. Such procedures are termed "naive" procedures. Let $\tilde{\lambda}_0 = (\tilde{\gamma}_0, 0)$ be the naive estimate under H_0 (solution of $\mathsf{S}(\lambda,\mathbf{w}) = 0$, under $H_0 : \theta = 0$). According to Valença (2003), two naive score statistics to test the homogeneity hypothesis can be defined:

$$Z_O = \frac{\frac{1}{2}\sum_{i=1}^{k} h_i(\tilde{\gamma}_0, w_i)}{\sqrt{V_O(\tilde{\lambda}_0,\mathbf{w})}}, \qquad \text{and} \qquad Z_H = \frac{\frac{1}{2}\sum_{i=1}^{k} h_i(\tilde{\gamma}_0, w_i)}{\sqrt{V_H(\tilde{\lambda}_0,\mathbf{w})}}, \tag{6}$$

where h_i is given in (5), with x_{ij} replaced by w_{ij} and $V_0(\tilde{\lambda}_0,\mathbf{w}) = I_{\theta\theta}(\tilde{\lambda}_0,\mathbf{w}) - I_{\theta\gamma}(\tilde{\lambda}_0,\mathbf{w})\left(I_{\gamma\gamma}(\tilde{\lambda}_0,\mathbf{w})\right)^{-1} I_{\gamma\theta}(\tilde{\lambda}_0,\mathbf{w})$, considering that $I_{\theta\theta}$, $I_{\theta\gamma}$, $I_{\gamma\theta}$ and $I_{\gamma\gamma}$ are elements of I, which is partitioned according to $\lambda = (\boldsymbol{\gamma}^{\mathsf{T}}, \theta)$. $V_H(\tilde{\lambda}_0,\mathbf{w}) = \frac{1}{4}\sum_{i=1}^{k}(h_i(\tilde{\gamma}_0, w_i) - \bar{h}(\tilde{\gamma}_0,\mathbf{w}))^2$, with $\bar{h} = \sum h_i/k$.
However, as is well known (Gimenez and Bolfarine, 2000), the naive score function $\mathsf{S}(\lambda,\mathbf{w})$ is biased and leads to inconsistent inference.

4 The Corrected Score Statistic

The corrected score approach for consistent inference in measurement error model was considered in Nakamura (1990) and Gimenez and Bolfarine (1997). The corrected score function $\mathsf{S}^*(\lambda;\mathbf{w}) = \mathsf{S}^*(\lambda;\mathbf{w},Y)$ is defined as a function whose conditional expectation $E\left[\mathsf{S}^*(\lambda;\mathbf{w},Y)|\mathbf{x},Y)\right]$ with respect to $\mathbf{w}$ given $(\mathbf{x},Y)$, coincides with $\mathsf{S}(\lambda;\mathbf{x})$. Under the normality assumption for the measurement error, properties of the normal moments generating function can be used to obtain the corrected score vector for the Weibull model (1). For this model, we can obtain a corrected log-likelihood function l^* so that $\mathsf{S}^*(\lambda;\mathbf{w}) = \partial l^*(\lambda,\mathbf{w})/\partial\lambda$, which is given by

$$\begin{aligned} l^*(\lambda,\mathbf{w}) &= l_0^*(\gamma,\mathbf{w}) \\ &\quad + \sum_{i=1}^{k} \log\left[1 + \frac{h_i^*(\gamma,w_i)\theta}{2} + \sum_{m=3}^{\infty} \frac{D_i^{*(m)}(\gamma,w_i)\theta^{\frac{m}{2}} E(V_i^m)}{m!}\right], \end{aligned}$$

where $$l_0^*(\gamma,\mathbf{w}) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left\{\delta_{ij}\left\{s(w_{ij}) - \log\sigma\right\} - \exp(s(w_{ij}) - f)\right\},$$

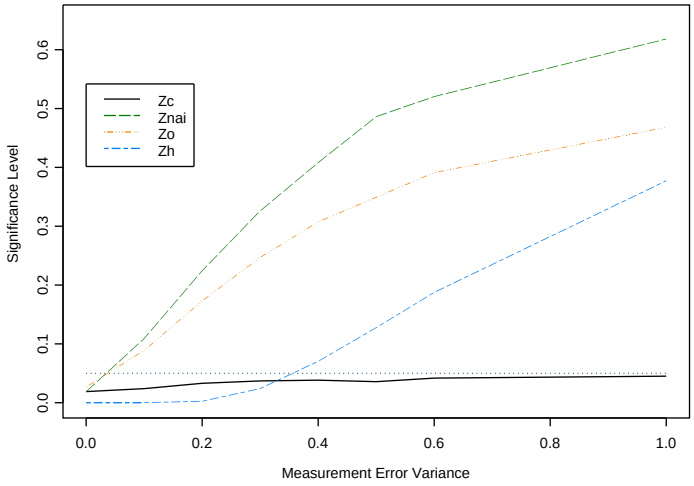


FIGURE 1. *Simulated levels of corrected* (Z_C) *and naive* (Z_0, Z_H, Z_{naive}) *tests of homogeneity for increasing values of variance of the measurement error (1,000 replications). Without Censoring.*

$$h_i^*(\gamma, w_i) = \frac{1}{\sigma^2}\left\{\left[\sum_{j=1}^{n_i}\left(e^{s(w_{ij})-f} - \delta_{ij}\right)\right]^2 - \sum_{j=1}^{n_i} e^{s(w_{ij})-f} - F_i\right\}, \quad (7)$$

with $F_i = \sum_{j=1}^{n_i}\left[\exp(2s(w_{ij}) - 2f) - \exp(2s(w_{ij}) - 4f)\right]$, $s(w_{ij}) = (y_{ij} - \alpha - \beta_z^T \mathbf{z}_{ij} - \beta_x w_{ij})/\sigma$ and $f = (\beta_x^2\phi)/2\sigma^2$. $D_i^{*(m)}(\gamma, w_i)$ is such that, $E[D_i^{*(m)}(\gamma, w_i)|Y, x_i] = D_i^{(m)}(\gamma, x_i)$, given in (4). Then, we can compute, under the null hypothesis, the corrected information matrix $I^*(\lambda_0, \mathbf{w}) = -(\partial \mathsf{S}^*(\lambda; \mathbf{w})/\partial\lambda)|_{\lambda=\lambda_0}$, where $\lambda_0 = (\lambda, 0)$. A proposed corrected score statistic for testing $H_0 : \theta = 0$, based on results given in Gimenez et al. (2000), is given by

$$Z_C = \frac{\frac{1}{2}\sum_{i=1}^{k} h_i^*(\hat{\gamma}^*, w_i)}{\sqrt{V_C(\hat{\lambda}_0^*, \mathbf{w})}}, \quad (8)$$

where $\hat{\lambda}_0^* = (\hat{\gamma}^*, 0)$, is the correct estimate under the null hypothesis (solution of $\mathsf{S}^*(\lambda; \mathbf{w}) = 0$, under $H_0 : \theta = 0$) and h_i^* is given in (7). Considering the matrices partitioned according to $\lambda = (\boldsymbol{\gamma}^{\mathsf{T}}, \theta)$, V_C is defined as

$$V_C(\lambda_0, \mathbf{w}) = \left[V_o^*(\lambda_0, \mathbf{w})\right]^2 G_{\theta\theta}^*(\lambda_0, \mathbf{w}),$$

where $V_o^*(\lambda_0, \mathbf{w}) = I_{\theta\theta}^*(\lambda_0, \mathbf{w}) - I_{\theta\gamma}^*(\widehat{\lambda}_0, \mathbf{w})\left\{I_{\gamma\gamma}^*(\lambda_0, \mathbf{w})\right\}^{-1} I_{\gamma\theta}^*(\lambda_0, \mathbf{w})$,

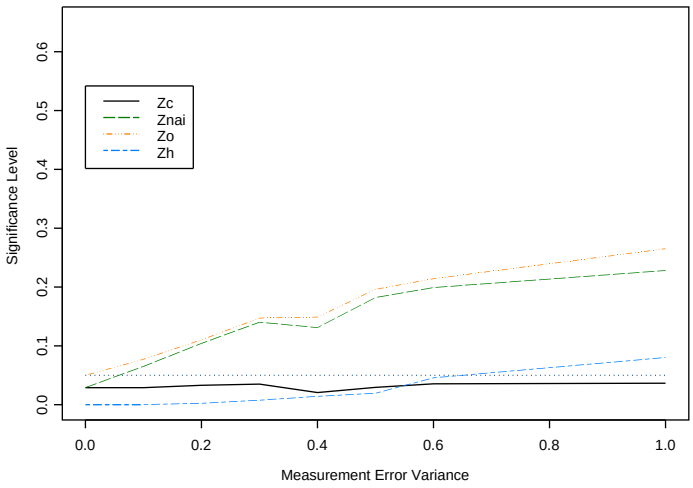


FIGURE 2. *Simulated levels of corrected (Z_C) and naive (Z_0, Z_H, Z_{naive}) tests of homogeneity for increasing values of variance of the measurement error (1,000 replications). 50% Censored.*

with $I^*_{\theta\theta}$, $I^*_{\theta\gamma}$, $I^*_{\gamma\theta}$ and $I^*_{\gamma\gamma}$ being elements of I^*. $G^*_{\theta\theta}$ is an element corresponding to θ, of the matrix $G^*(\lambda_0, \mathbf{w}) = I^{*-1}(\lambda_0, \mathbf{w})\widehat{\Gamma}^*(\lambda_0, \mathbf{w})I^{*-1}(\lambda_0, \mathbf{w})$, with $\widehat{\Gamma}^*(\lambda_0, \mathbf{w}) = \sum_{i=1}^{k} \mathsf{S}_i^*(\lambda_0, \mathbf{w}_i)\mathsf{S}_i^{*\mathsf{T}}(\lambda_0, \mathbf{w}_i)$.
We consider that to a large number of groups the distribution of Z_C, under H_0, can be approximate to a standard normal distribution, and we can use the statistic to performing a unilateral test which reject the null hypothesis to large positive values of this statistic.
We can use the sandwich structure of the variance of Z_C to define another naive statistic, $Z_{naive} = \frac{1}{2}\sum_{i=1}^{k} h_i(\tilde{\gamma}_0, w_i)/(V_{naive}(\tilde{\lambda}_0, \mathbf{w}))^{1/2}$, where $\tilde{\lambda}_0 = (\tilde{\gamma}_0, 0)$ is the naive estimate under H_0, $h_i(\tilde{\gamma}_0, w_i)$ is given in (5), with x_{ij} replaced by w_{ij} and V_{naive} is obtained similarly to V_C, but using the usual observed information matrix and the sandwich estimator instead of the corrected functions.

5 Simulation

A simulation study was conducted to compare the performance of the proposed test based on the corrected score Z_C, with the naive tests Z_O and Z_H given in (6) and Z_{naive}. We use a sample with $k = 100$ and $n_i = 5$ and test the homogeneity hypothesis. We carried out size simulations based on 1,000 replications, considering increasing values for the variance of measurement

error (ϕ), for a nominal size of 5%. We consider uncensored samples and samples with 50% of censoring. The response was generated for z_{ij} scalar, assuming that $\alpha = 0.5$, $\beta_z = 0.8$, $\beta_x = 1$, and $\sigma = 0.75$. The random variables V_i were generated as *i.i.d.* $N(0, 1)$. The results are reported in Figures 1 and 2. It is clear that the naive tests get simulated sizes very far from the nominal levels with the increasing of ϕ, while the corrected test Z_C has simulated sizes close to the nominal level. Although censoring has the effect of reduce the level of the test for the naive tests, in general the use of the corrected test Z_C leads to improvement in the level of the test.

References

Gimenez, P. and Bolfarine, H. (1997). Corrected score functions in classical error-in-variables and incidental parameter models. *Australian Journal of Statistics*, **39**, 325–344.

Gimenez, P., Bolfarine, H., and Colosimo, E. (1999). Estimation in Weibull regression models with measurement errors. *Communication in Statistics, Theory and Methods*, **28**(2), 495–510.

Gimenez, P., Bolfarine, H., and Colosimo, E. (2000). Hypotheses testing in measurement error models. *Annals of the Institute of Statistical Mathematics*, **52**(4), 698–711.

Liang, K.Y. (1987). A locally most powerful test for homogeneity with many strata. *Biometrika*, **74**, 259–64.

Nakamura, T. (1990). Corrected score functions for error-in-variables models: methodology and applications to generalized linear models. *Biometrika*, **77**, 127–137.

Valença, D.M. (2003). *Testes de homogeneidade e estimação para dados de sobrevivência agrupados e com erros de medida*. Doctoral thesis. Department of Statistics. University of São Paulo.

Structural Accelerated Failure Time versus Proportional Hazards Modelling for the Effects of Observed Exposures on Repeated Events in a Clinical Trial

An Vandebosch[1], Els Goetghebeur[1], and Lut Van Damme[2]

[1] Dept. of Applied Mathematics and Computer Science, Ghent University. Krijgslaan 281- S9, B-9000 Ghent, Belgium.
[2] Conrad Program, 1611 North Kent Street, Suite 806, Arlington, VA 22209, USA.

Abstract: We consider causal inference for randomised clinical trials which assign patients to an experimental treatment or control and observe repeated event times as outcomes. The structural accelerated failure time models (Robins and Tsiatis, 1991) allow to analyze causal effects of observed exposures on a single non-informatively censored survival time. They parametrize an exposure-specific transformation of the observed outcome into a potential exposure-free outcome. Estimation relies on equality of exposure-free distributions between treatment arms and requires parameter-dependent recensoring of events.
As an alternative, Loeys and Goetghebeur (2002) proposed structural Cox PH models for the analysis of possibly selective, time-constant exposures. They transform observed hazard rates rather than observed times into potential exposure-free hazard rates via a function of observed exposure. They thus avoid recensoring. To handle repeated survival times we extend both structural methods following the marginal modelling principle. We discuss model-specific constraints and compare advantages. We apply both approaches to analyze recurrent lesions in an HIV-prevention trial.

Keywords: Accelerated failure time models; Causal inference; Proportional hazards model.

1 Introduction

We consider randomised clinical trials which assign patients to an experimental treatment or control and observe repeated events as outcomes. Once a significant intent-to-treat effect is found one aims to quantify a causal dose-response relationship for the experimental treatment. This must account for treatment actually received when varying levels of exposure to the prescribed dosing regimen are seen. An observed association between dose and response does not just reflect a causal effect when different exposure levels correspond with different risks even in the absence of a treatment

effect. To answer the causal question: "what if experimental exposures had been withheld?" potential outcomes are introduced and assumptions for identifiability are imposed on the latent variables.
Structural Accelerated Failure Time(SAFT) models (Robins and Tsiatis, 1991) parametrize the transformation of observed times into potential exposure-free times in function of observed exposure. Estimation equations impose equally distributed exposure-free survival times between randomized arms. To avoid informative censoring on the backtransformed scale, a recensoring scheme is typically devised which can dramatically reduce the number of events in the backtransformed dataset and thus reduces the information available for causal inference.
As an alternative to structural AFT models, structural Cox PH models were recently proposed by Loeys and Goetghebeur (2002) to analyze the effect of possibly selective but time-averaged exposures in randomised clinical trials. Rather than transforming survival times directly, they transform observed hazard rates into potential exposure-free hazard rates via a function of observed exposure. In doing so they avoid the need for recensoring.
To analyze events occurring repeatedly over time, we extend both the SAFT model and the structural PH model. Specifically, we adapt the marginal methodology of Wei, Lin and Weissfeld (1989) and use robust variance estimators to correct for the possible correlations within women. We compare advantages and constraints of both methods and apply them to analyze data from an HIV prevention trial which we describe next.

2 A Causal Question in the COL-1492 Trial

In the randomised, placebo-controlled COL-1492 trial for the prevention of HIV (Van Damme *et al*, 2002), female sex workers from different centres in Africa and Asia were triple blindly randomised to either an experimental vaginal gel, COL-1492, or a placebo gel, Replens. On both arms of the trial, women were asked to use the assigned gel before every vaginal act and the male condom for every sexual act. At monthly scheduled clinic visits they were tested for HIV, sexually transmitted infections and lesions. The primary intent-to-treat analysis revealed a significantly negative effect of the assigned experimental treatment on HIV-incidence.
The hypothesis was raised that high gel exposure might cause lesions which ultimately give the virus easier access. At every clinic visit women reported on their sexual acts of various types and the preventive measures that had actually been taken. Here we set out to estimate the causal effect of the daily number of vaginal acts with experimental gel on the incidence of lesions, a recurrent event.
The largest centre Durban had the highest retention rate in the study (93% after 6 months), saw a borderline significant HIV-effect (Hazard Ratio = 1.6, p=0.06) and the largest number of observed lesions. We therefore

concentrate our efforts on estimating the causal effect of gel exposure on lesions in that centre.

3 A Structural AFT Approach for Recurrent Lesions

Let $R_i = 1(0), i = 1, \cdots, n$, denote whether the i^{th} person was assigned to the treatment (control) arm. Consider further the times to K successive events (observed lesions) $T_{1i}, \ldots, T_{Ki}$, which may be noninformatively censored by C_i (end of study participation). Let E_{ki} be the daily number of vaginal acts with the experimental gel between events $k-1$ and k (where event 0 means admission into the trial). Hence $E_{ki} = 0, \forall k$ for a woman on the control arm. We propose a SAFT-model that backtransforms observed failure times T_{ki} to potential exposure-free failure times T^0_{ki} which would have been observed on the Replens arm. Specifically we assume that for some value β_0 of β:

$$T_{ki}(\beta) \stackrel{d}{=} \sum_{l=1}^{k} (T_{li} - T_{(l-1)i}) e^{-\beta E_{li}} \tag{1}$$

where $T_{0i} \equiv 0$ and $\stackrel{d}{=}$ means equality in distribution. At the true value β_0, $T_{ki}(\beta_0) \equiv T^0_{ki} \perp\!\!\!\perp R_i$ holds.

To transform censoring times we define a new censoring variable $C_{ki}(\beta) = \min(C_i, C_i e^{-\beta m_k})$, where $m_k = \max_{i=1}^{n} \{E_{1i}, \ldots, E_{ki}\}$. Now testing for $H_0 : \beta_0 = \beta$ amounts to testing for independence between $T_{ki}(\beta)$ and R_i based on noninformatively right-censored pseudo-data $\min(T_{ki}(\beta), C_{ki}(\beta))$ with censoring indicators $I(T_{ki}(\beta) \leq C_{ki}(\beta))$. We use the robust score test from a working Cox PH model with a common effect of treatment for the distinct event times $T_{ki}(\beta)$ but possibly different baseline risks. The estimated causal effect $\hat{\beta}_0$ is then the value of β which minimizes this test statistic. A $(1-\alpha)100\%$ confidence interval becomes the set of values of β which are not rejected by this robust score test at the α significance level. When analyzing T_{1i}, time to first event, in the COL-1492 trial we obtain $e^{-\hat{\beta}_0} = 1.68$ with 95% CI [1.00, 3.13]. As illustrated in Figure 1, the marginal SAFT-model which includes all $K = 6$ possible failure times, leads to an estimated effect $e^{-\hat{\beta}_0}$ of 1.93 with 95% CI [1.06, 3.13]. Thus for women observed with one treated act per day, the k^{th} lesion time is estimated to be 1.93 times longer under the potential Replens regimen.

The recensoring scheme explained above reduces an original 62 observed events to just 35 in the pseudo data set at the estimated causal effect; while only 25(54) at the lower (upper) 95% confidence limit remain. More sophisticated and economical recensoring schemes could however be devised. This wins in importance as more structural parameters enter the model. To avoid recensoring we explore a structural PH model for the recurrent events, which transforms estimated intensity functions rather than survival times themselves.

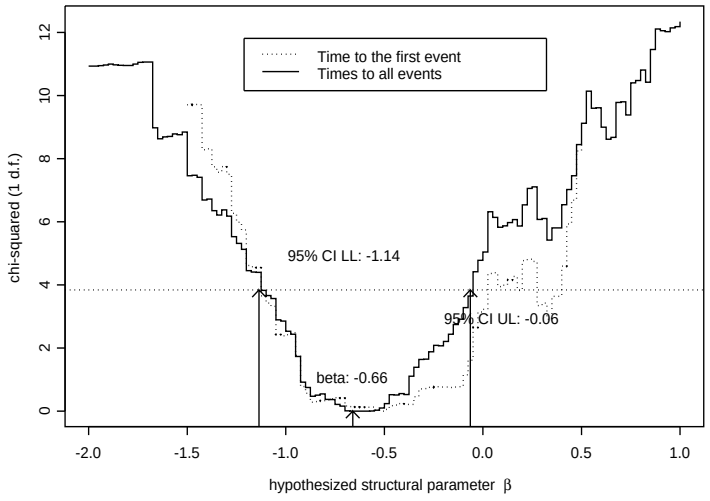


FIGURE 1. *Value of the robust score test statistic as a stepfunction of β with the estimated causal effect $\hat{\beta}_0$ and its 95% confidence limits.*

4 A Structural PH Approach for Recurrent Lesions

Per woman we consider now a time-averaged measure of exposure, E_i, over the observation period. In addition we define U_i^E the potential exposure to the experimental gel which is experienced when person i is assigned to it. On the experimental arm $U_i^E = E_i$ but in the control arm it is unobserved. By design U_i^E is equally distributed between randomised arms.
The structural PH model then transforms the compliance-specific hazard rate of the k^{th}-failure in the experimental arm, $\lambda_k(t|U_i^E = u, R_i = 1)$ to its counterpart in the placebo arm $\lambda_k(t|U_i^E = u, R_i = 0)$ via a function of observed exposure and the causal effect ψ_0:

$$\lambda_k(t|U_i^E = u, R_i = 0) = \lambda_k(t|U_i^E = u, R_i = 1)e^{-\psi_0 u} \tag{2}$$

Because U_i^E is latent in the control arm the hazard rate on the left hand side is not directly estimable .
To estimate ψ_0, we will compare estimated hazards between randomised arms. At the true value, the corresponding survival distributions must be equal when model (2) holds. More specifically, we estimate the failure-specific survival distribution in the control arm: $S_k(t|R_i = 0)$ by the

Kaplan-Meier method. For a given ψ, we compute the exposure-free failure-specific survival distributions in the treatment arm, following: $S_{k,1\to 0}(t|\psi) = \frac{\sum_{i=1}^{n} R_i S_k(t|R_i=1,U_i^E)^{exp\left\{-\psi U_i^E\right\}}}{\sum_{i=1}^{n} R_i}$

When model (2) holds together with the randomisation assumption, it implies that

$$S_k(t|R_i = 0) = S_{k,1\to 0}(t|\psi_0); \forall k.$$

A stratified 'class K' test (Gill, 1980) with an empirically estimated variance is devised to compare those estimated distribution functions between arms. The value of ψ which minimizes this test statistic yields again the estimated causal effect.

Conclusion

We find that both the structural AFT and structural PH approaches are adaptable to the marginal analysis of recurrent event times. In their current implementation, the former has the advantage of being able to handle time-varying exposures while the latter avoids efficiency loss due to recensoring. To guide the practical choice between them one must currently consider model fit, impact of recensoring and the variation of exposure rates over time.

References

Gill, R.D. (1980). *Censoring and Stochastic Integrals.* Mathematical Centre Tracts 124: Amsterdam.

Loeys, T. and Goetghebeur, E. (2003). A causal proportional hazards estimator for the effect of treatment actually received in a randomized trial with all-or-nothing compliance. *Biometrics*, **59**(1), 100–105.

Loeys, T. and Goetghebeur, E. (2002). Causal proportional hazards models for the effect of treatment actually received in a randomized trial with selective noncompliance. *Technical Report 2002, Ghent University.*

Robins, J.M. and Tsiatis, A.A. (1991). Correcting for non-compliance in randomised trials using rank preserving structural failure time models. *Communications in statistics: Theory and Methods*, **20**(8), 2609–2631.

Van Damme, L. et al (2002). Effectiveness of COL-1492, a nonoxynol-9 vaginal gel, on HIV-transmission among female sex workers. *The Lancet*, **360**, 971–977 .

Wei, L.J., Lin, D.Y., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, **84**(408), 1065–1073.

Modelling Pasture Growth Rates Using L-spline Mixed Models

S.J. Welham[1], B.R. Cullis[2], G. Li[2], M.G. Kenward[1], and R. Thompson[3]

[1] London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK
[2] Wagga Wagga Agricultural Institute, NSW 2650, Australia
[3] Rothamsted Research Ltd, Harpenden AL5 2JQ, UK

Abstract: L-splines use a linear differential operator to define an underlying form for a model, then fit a smooth curve subject to a penalty defined by the linear differential operator and a smoothing parameter. This paper describes briefly the use of L-splines within the mixed model framework, with application to the estimation of pasture growth rates in a complex long-term rotation experiment.

Keywords: L-splines; Mixed models; REML; Smoothing splines.

1 Introduction

This paper describes the use of mixed model L-splines to analyse a long-term rotation experiment in order to predict pasture growth rates. Modelling of a pair of smooth curves for each rotation is required to estimate pasture growth rate. As the curves do not correspond to a parametric form, smoothing splines provide a convenient method of fitting curves.
Cubic splines were introduced into the mixed model setting by Wang (1998), Brumback and Rice (1998) and Verbyla *et. al.*, (1999), who noted the mathematical equivalence between the penalised sum of squares used to fit a cubic spline and the mixed model equations for a specific mixed model. The smoothing parameter can then be estimated as part of the model fitting process, usually by REML. In this context, the cubic spline can be built into a general treatment structure and fitted at different levels of the structure. In addition, random terms can easily be added to account for other sources of variation in the data.
In the mixed model context, the cubic spline is partitioned into a fixed linear component plus random terms, with zero expectation, representing smooth deviations about the linear trend. The penalty is expressed in terms of the accumulated second derivative of the fitted curve. However, in the rotation experiment the underlying trend is not linear, but a seasonal cycle with some linear trend. In this case the cubic spline model seems unnatural, as it is clear that the deviation about the linear trend is non-zero. In this

case it seems more natural to use L-splines, which use an underlying form defined by a linear differential operator and penalise departures from this underlying form. L-splines based on trigonometric functions can be fit as linear mixed models. In this paper, we use an L-spline with underlying form based on linear trend plus simple periodic cycles to model the data from the rotation experiment.

2 Rotation Experiment

The data analysed in this paper come from a large rotation experiment designed to investigate the influence of lime application and different rotations on pasture growth rate. Further details of the experimental approach and design can be found in Li *et. al.*, (2001). The experimental design consisted of 2 blocks of 40 plots, with eight treatments, using a 2^3 factorial structure with underlying factors for perennial versus annual pastures, continuous pastures versus pasture with crop rotations, and limed versus unlimed.
During any season, each block contained 2, 3 or 6 replicate plots in pasture for each treatment. For grazing, a three-plot rotation system of replicate plots was used for all treatments except for the annual pasture/crop rotation treatment where a two-plot rotation system was used. The length of spell was therefore effectively nested within the rotations. Measurements of available pasture were taken from plots when a grazing period ended (after grazing) and at the end of the spell just before the next grazing period (before grazing). Five years data were available.
The aim of the analysis was to assess differences in the rate of production of dry matter for the eight treatments, measured as the rate of dry matter production during the spell between grazing. We used an indirect approach to estimating relative growth rates via modelling the log of the dry matter measurements as two paired series of measurements: after grazing and before grazing. The shape of these separate curves could be modelled closely in time using splines. For a constant grazing cycle with spell length s and stocking rate, the relative growth rate at any time can then be estimated from the lagged difference between the two curves, using the spell time as the lag. Disturbances in the grazing cycle were adjusted for empirically by using lack of fit terms that account for variation in measurements around the underlying smooth curves. In addition, we directly model correlation over time due to repeated measurements from each plot.

3 L-splines

In the simplest case, we have data $\mathbf{y}$ $(n \times 1)$, measured with error, and explanatory variable $\mathbf{x}$ with n unique values $\{x_1, x_2 \ldots x_n\}$. We specify an underlying form of curve that is appropriate to the data, described by a set of core functions $u = \{u_j; j = 1 \ldots m\}$ which are annihilated by a linear

differential operator L, *i.e.* $Lu = 0$. An L-spline penalises departures from the favoured form defined by $Lu = 0$. For a single spline term, we fit a model of the form

$$\mathbf{y} = \mathbf{g} + \mathbf{e}$$

where $\mathbf{g} = g(\mathbf{x})$ for a function $g(t)$. The L-spline is the function $g(t)$ that minimises the penalised sum of squares

$$(\mathbf{y} - \mathbf{g})'(\mathbf{y} - \mathbf{g}) + \lambda \int [Lg(s)]^2 ds \tag{1}$$

where λ is a parameter that controls the amount of smoothing. Ramsay and Silverman (1997, section 15.2) prove that, for any basis $\{u_j\}$ such that $Lu_j = 0$ for $j = 1 \ldots m$, the function g minimising the penalised sum of squares (for given λ) has the form

$$g(t) = \sum_{j=1}^{m} d_j u_j(t) + \sum_{i=1}^{n} c_i k_2(x_i, t)$$

where k_2 is a reproducing kernel function. Recipes for constructing a set of basis functions for an L-spline with initial value constraints are given by Heckman & Ramsay (2000). Properties of the reproducing kernel function can be used to show that the penalised sum of squares (1) can be written as

$$(\mathbf{y} - \mathbf{U}\mathbf{d} - \mathbf{K}\mathbf{c})^T(\mathbf{y} - \mathbf{U}\mathbf{d} - \mathbf{K}\mathbf{c}) + \lambda \mathbf{c}^T \mathbf{K}\mathbf{c} \tag{2}$$

where the matrices $\mathbf{U}$ and $\mathbf{K}$ have elements $[\mathbf{U}]_{ij} = u_j(x_i)$ for $i = 1 \ldots n, j = 1 \ldots m$, $[\mathbf{K}]_{ij} = k_2(x_j, x_i)$ for $i, j = 1 \ldots n$, $\mathbf{d}$ is a vector of length m and $\mathbf{c}$ is a vector of length n.

4 L-spline Mixed Models

Minimisation of the penalised sum of squares (2) requires solution of the equations

$$\begin{bmatrix} \mathbf{U}'\mathbf{U} & \mathbf{U}'\mathbf{K} \\ \mathbf{K}'\mathbf{U} & \mathbf{K}'\mathbf{K} + \lambda\mathbf{K} \end{bmatrix} \begin{pmatrix} \hat{\mathbf{d}} \\ \tilde{\mathbf{c}} \end{pmatrix} = \begin{bmatrix} \mathbf{U}'\mathbf{y} \\ \mathbf{K}'\mathbf{y} \end{bmatrix}$$

This system of mixed model equations contains implicit constraints $\mathbf{U}'\tilde{\mathbf{c}} = 0$. We can absorb these constraints into the equations by reparameterising the penalised sum of squares in terms of δ (size $n - m$) where $\mathbf{c} = \mathbf{C}\delta$ for any $n \times n - m$ matrix $\mathbf{C}$ of full column rank such that $\mathbf{U}'\mathbf{C} = 0$. The resulting amended equations correspond to the mixed model equations for a linear mixed model of the form

$$\mathbf{y} = \mathbf{U}\mathbf{d} + \mathbf{K}\mathbf{C}\delta + \mathbf{e}$$

with $\mathbf{U}\mathbf{d}$ representing fixed terms in the model, $\mathbf{K}\mathbf{C}\delta$ representing a random model term, and a residual vector $\mathbf{e}$ with $\text{var}(\mathbf{e}) = \sigma^2\mathbf{I}$, $\text{var}(\delta) = \sigma_s^2\mathbf{H}^{-1}$ and

$\lambda = \sigma^2/\sigma_s^2$, where $\mathbf{H} = \mathbf{C}'\mathbf{K}\mathbf{C}$. As for cubic spline mixed models (Verbyla *et. al.*, 1999, Wang 1998, Brumback and Rice, 1998), the spline can then be estimated by fitting this mixed model with the smoothing parameter estimated as a variance parameter using REML. For data sets with many distinct covariate values, the L-spline mixed model defined above can be difficult to fit, as it may require a large amount of workspace and processing time. Both these factors can be reduced by using a reduced number of knots, r say, defined at distinct covariate values $\mathbf{t} = (t_1,\ t_2 \ldots t_r)'$.

5 Analysis of Rotation Experiment

The L-spline with core functions $\{1,\ t,\ \sin(\omega t),\ \cos(\omega t)\}$ was used in modelling the log dry matter data. The basic treatment structure of the experiment at any one time can be written as $rotation/graztrt$ where $rotation$ is a factor describing the 8 different rotations and $graztrt$ describes the status of each plot with respect to its position in the grazing pattern (after/before). This leads to the following full model (using Genstat/ASREML operators)

$$\begin{aligned} fixed &\sim (constant + lin(t) + \cos(\omega t) + \sin(\omega t)) * (graztrt/rotation) \\ random &\sim plot + lspl(t) + fac(t) + rotation.lspl(t) + rotation.fac(t) \\ &\quad + graztrt.rotation.lspl(t) + graztrt.rotation.fac(t) \\ residual &\sim plot.power(t) \end{aligned} \tag{3}$$

where t is the number of days since 1 April 1992 and $\omega = 2\pi/365.25$ is used to give a periodic cycle with period 1 year (on average). The continuous variables $lin(t)$, $\cos(\omega t)$ and $\sin(\omega t)$ have the obvious definition. The spline function $lspl(t)$ represents L-spline basis functions using 12 equally-spaced knot points per year. The term $fac(t)$ represents a factor version of the time variable t and fits a separate effect for each distinct value of t present in the data. Lack of fit terms (ie. terms including $fac(t)$) are used to account for variation in the grazing pattern. The residual term fits a common power model within plots to account for correlation in measurements across time within plots. The analysis was done using GenStat and ASREML.
The estimated daily relative rate of pasture growth at time t can then be calculated as

$$\hat{f}(t + \frac{s}{2}) = \frac{\hat{B}(t+s) - \hat{A}(t)}{s}$$

where s is the spell time and $B(t), A(t)$ represent the log dry matter data at time t before/after grazing respectively. Integration across time can be used to estimate pasture accumulation. As these quantities are all linear functions of model parameters, it is straight-forward to obtain their predicted values with standard errors.

6 Discussion

We have illustrated the use of L-spline mixed models in the analysis of complex experimental data where the underlying form has a periodic pattern. This work adds to the growing list of types of smoothing splines that can be fitted within mixed models. Other examples include the P-splines of Eilers and Marx (1996) and the related penalised splines of Parise et al (2001). Further work is required to determine the relative merits of different spline types in the mixed model context.

Acknowledgments: Several authors (SW, GL, BC) would like to acknowledge the funding of the Grains Research and Development Corporation of Australia.

References

Brumback, B.A. and Rice, J.A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *Journal of American Statistical Association*, **93**, 961–994.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89–112.

Heckman, N.E. and Ramsay, J.O. (2000). Penalized regression with model-based penalties. *Canadian Journal of Statistics*, **28**, 241–258.

Li, G.D., Helyar, K.R., Conyers, M.K., Cullis, B.R., Cregan P.D., Fisher, R.P., Castleman, L.J.C., Poile, G.J., Evans, C.M., and Braysher, B. (2001). Crop responses to lime in long-term pasture-crop rotations in a high rainfall area in south-eastern Australia. *Australian Journal of Agricultural Research*, **52**, 329–341.

Parise, H., Wand, M.P., Ruppert, D., and Ryan, L. (2001). Incorporation of historical controls using semiparametric mixed Models. *Applied Statistics*, **50**, 31–42.

Ramsay, J.O. and Silverman, B.W. (1997). *Functional Data Analysis.* New York: Springer-Verlag.

Verbyla, A.R., Cullis, B.R., Kenward, M.G., and Welham, S.J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *Applied Statistics*, **48**, 269–311.

Wang, Y. (1998). Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society, Series B*, **60**, 159–174.

A Model Based View of Partial Least Squares

Joe Whittaker

[1] Statistics Dept., Lancaster University, LA1 4YF, UK.
Email: joe.whittaker@lancaster.ac.uk

Abstract: Electronic instrumentation provides many examples of 'fat' data ($n<p$) and one succesful method, PLS, for its analysis is somewhat neglected by statisticians. We show how tracking the independence graphs of the sequence of transformations of the PLS algorithm leads to a clear description of the independence structure of the derived components and to formulating a bilinear factor model for generating the observed data.

Keywords: Graphical model; Prediction; Partial least squares.

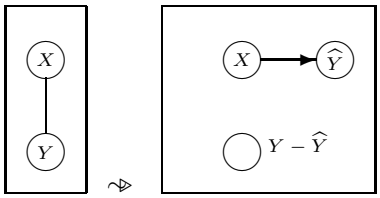
1 Introduction

PLS originates from some papers of Wold in the 60's and 70's, and, while there is no universal agreement on the best description of the PLS procedure, if the variance matrix of the explanatory variables X is singular, it is acknowledged that PLS performs well when ordinary least squares fails. PLS has many adherents, especially from chemometrics and food science; it is a technique that works well in practise, especially for the analysis of data incorporating spectrometer readings. Martens and Naes (1989) give an extensive practical exposition. An excellent recent theoretical review of the statistical basis for PLS is given by Helland (2001).
The statistical community is somewhat more sceptical. Stone and Brooks (1990) suggests that the methodology, in particular the choice of optimisation criterion, is arbitrary. Later Butler and Denham (2000) point out some rather non-intuitive shrinkage properties of the PLS procedure. However in applications, it is empirically competitive with other similar statistical procedures such as ridge regression and principal components regression, for instance see Frank and Friedman (1993).
In Section 2 we summarise the linear least squares prediction framework needed to build independence graphs for the PLS procedure which itself is described in Section 3. The paper ends with a few concluding remarks.

2 Linear Least Squares Prediction

We follow Whittaker (1990) but also see Christensen (1991). Suppose that $X(p \times 1)$ and $Y(q \times 1)$ are random vectors with some joint distri-

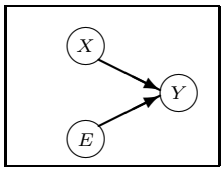
bution, and with expectations $\mathsf{E}(X)$ and $\mathsf{E}Y$. The covariance $\mathsf{cov}(X, Y)$ is a bilinear operator, and for the purposes of this paper we suppose that $X \perp\!\!\!\perp Y \iff \mathsf{cov}(X, Y) = 0$ identifying independence with zero covariance. The LLSP is $\mathsf{E}(Y|X) = \mathsf{E}(Y) + \mathsf{cov}(Y, X)\mathsf{var}(X)^{-1}[X - \mathsf{E}(X)]$. The alternative shorthand notations for the LLSP, $\widehat{Y}$ and $\widehat{Y}(X)$ prove useful. Note that $\widehat{Y} = BX$ where $B = \mathsf{cov}(Y, X)\mathsf{var}(X)^{-1}$ is the matrix of prediction coefficients. The LLSP satisfies the important orthogonality condition $\mathsf{cov}(Y - \mathsf{E}(Y|X), X) = 0$ or, equivalently, $Y - \mathsf{E}(Y|X) \perp\!\!\!\perp X$. The partial variance is defined as variance of the residual $\mathsf{var}(Y|X) = \mathsf{var}[Y - \mathsf{E}(Y|X)]$, and the partial covariance by $\mathsf{cov}(Y, Z|X) = \mathsf{cov}[Y - \mathsf{E}(Y|X), Z - \mathsf{E}(Z|X)]$. The adjective 'partial' has the same meaning as in partial correlation, rather than as in PLS. The criterion for conditional independence, Lauritzen (1996) or Whittaker (1990), is defined here by $X \perp\!\!\!\perp Y|Z \iff \mathsf{cov}(X, Y|Z) = 0$, and defines the missing edges in the independence graphs. The transformation: $(X, Y) \rightsquigarrow\!\!\!\!\blacktriangleright (X, Y - \widehat{Y})$, of the joint vector (X, Y) to X and residual of Y from the LLSP has associated graphs



The graph on the right displays the fact that $X \perp\!\!\!\perp Y - \widehat{Y}$, and that the link from X to $\widehat{Y}$ is deterministic. Thick arrows are used to display deterministic (logical or functional) relationships.
The decomposition

$$Y = \hat{Y} + Y - \hat{Y} = BX + Y - \hat{Y}$$

is a deterministic identity.
A generative statistical model can be built from this LLSP identity by supposing the random vector E exists exogenously, $E \perp\!\!\!\perp X$, B fixed, and $Y = BX + E$, with graph



If $\mathsf{var}(E) = \mathsf{var}(Y|X)$ and B is identical to the above prediction coefficients then samples of (X, Y) taken from the original distribution of (X, Y) or taken from (X, E) cannot be distinguished by inspection of second order statistical properties.

3 The PLS Population Algorithm

The PLS procedure is usually described by an algorithm, for instance, Martens and Naes (1989) pl19. We follow this but express the procedure entirely in terms of prediction of random vectors suppressing any reference to samples of size n, that is, in population terms, Helland (2001).
We suppose Y is scalar, $q = 1$, though this is easily generalised. The algorithm is

Initialise by setting $\mathsf{E}(X_i) = 0$ and $\mathsf{var}(X_i) = 1$ for $i = 1, \ldots, p$.

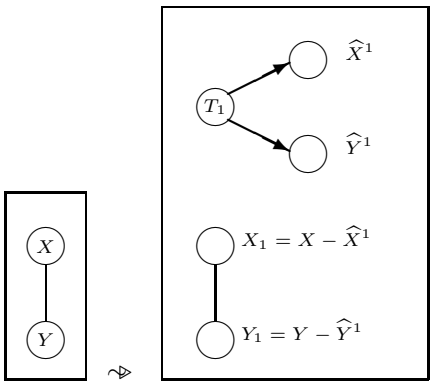
Repeat the two steps: 1 Compute the direction

$$c = \arg \max_{a|a'a=1} \mathsf{cov}(a'X, Y).$$

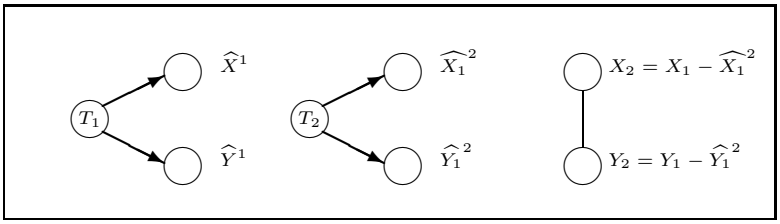
2 Adjust for $T = c'X$ by replacing X and Y by their residuals $X - \mathsf{E}(X|T)$ and $Y - \mathsf{E}(Y|T)$.

Stop the procedure at step k if $\mathsf{cov}(X, Y|T_1, \ldots, T_{k+1}) = 0$, resulting in components $(T_1, \ldots, T_k)$ from which the final predictor $\mathsf{E}(Y|T_1, \ldots, T_k)$ may be formed.

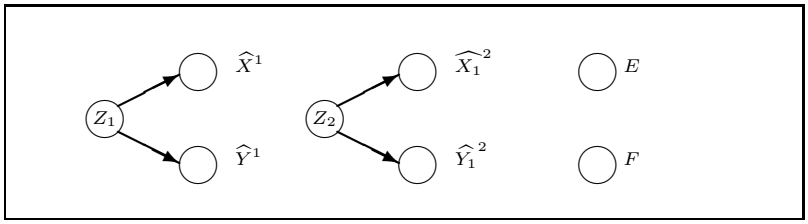
There are some obvious remarks to make about the procedure. It reduces the effective dimension of the explanatory model from p to k components. Any computation of $\mathsf{var}(X)^{-1}$ is avoided, and, the canonical correlation constraint $a'\mathsf{var}(X)a = 1$ is replaced by $a'a = 1$. These components are inherently informative about Y because of the way the loadings coefficients c are chosen, and so likely to be predictive. The components are mutually independent, $\perp\!\!\!\perp(T_1, T_2, \ldots, T_k)$ so that $\mathsf{E}(Y|T_1, T_2, \ldots, T_k)$ may be formed as a sum. There is a downside, the initial scaling is arbitrary as is the optimised criterion $\max_{a|a'a=1} \mathsf{cov}(a'X, Y)$. More importantly, there is no explicit statistical model, and thereby no default statistical inference.
The independence graphs that map this sequence of PLS transformations begin with the display of the graph for (X, Y) and, by extracting c_1 from (X, Y), introduce T_1 and the residuals (X_1, Y_1) after adjusting for T_1



Note that the component is independent of the residuals, $T_1 \perp\!\!\!\perp (X_1, Y_1)$, after adjusting for the the first component.
Extracting c_2 from the distribution of (X_1, Y_1) creates T_2.



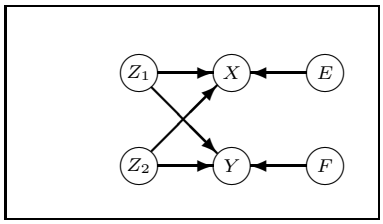
The procedure would stop at $k = 2$ if $X_2 \perp\!\!\!\perp Y_2$ or equivalently if $\mathsf{cov}(X_2, Y_2) = \mathsf{cov}(X, Y | T_1, T_2) = 0$, and there would be no edge between X_2 and Y_2 in this graph.
A generative model is formed by replacing the residuals by independent stochastic errors, E, F, as with the LS model above, by replacing the components by exogenous latent random variables, Z_1, Z_2, and by re-interpreting $\widehat{X}^1, \widehat{Y}^1$ as predictors from Z_1 rather than from T_1. Assuming $\perp\!\!\!\perp \{Z_1, Z_2, E, F\}$ their graph is



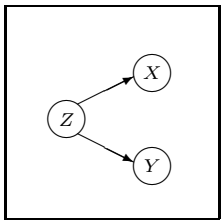
The observed values of X and Y are linear combinations of the predictors and the errors so that

$$\begin{aligned} X &= AZ + E \\ Y &= BZ + F \end{aligned}$$

and two versions of the independence graph for the generative model are



and



4 Concluding Remarks

This model is an instance of a common factor model, which has been described in the PLS literature by Martens and Naes (1989) and proposed for explicit model fitting by Burnham *et al.* (1999). The model as stated here is under specified. For instance, while the independence structure (zero covariances) is clear the variance matrices of the stochastic errors may be set in a variety of ways and additional assumptions are needed to make a full statistical data analysis.
Furthermore the final model is not specific to PLS. Any procedure that extracts sequences of linear combinations from (X, Y) and then from their residuals has the same sequence of independence graphs. The PLS choice, characterised by restriction to combinations of X alone with coefficients proportional to individual covariances, is not explicitly displayed in the graphs. For further insight, see Helland (2001).

References

Burnham, A.J. MacGregor, J.F., and Viveros, R. (1999). A statistical framework for multivariate latent variable regression methods based on maximum likelihood. *Journal of Chemometrics,* **13**, 49–65.

Butler, N.A. and Denham, M.C. (2000). The peculiar shrinkage properties of partial least squares regression. *Journal of the Royal Statistical Society, Series B*, **62**, 585–593.

Christensen, R. (1991). *Linear Models for Multivariate, Time series, and Spatial data.* New York: Springer.

Frank, I.E. and Friedman, J.H. (1993). A statistical view of some chemometrics regression tools. *Technometrics,* **35**, 109-135.

Helland, I.S. (1988). On the structure of partial least squares regression. *Communications in Statistics, Simulation and Computation*, **17**, 581–607.

Helland, I.S. (2001). Some theoretical aspects of partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, **58**, 97–107.

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

Martens, H. and Naes, T. (1989). *Multivariate Calibration.* Chichester: Wiley.

Stone, M. and Brooks, R.J. (1990). Continuum regression. *Journal of the Royal Statistical Society, Series B*, **52**, 237–269.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics.* Chichester: Wiley.

Wold, H. (1973). Non-linear iterative partial least squares (NIPALS) modelling: some current developments. In *Multivariate Analysis III* Ed. P.R. Krishnaiah, New York Academic Press. pp383-407.

Randomized Clinical Trials with a Pre- and Post-treatment Measure: Repeated Measures or ANCOVA?

Bjorn Winkens[1], Hubert J.A. Schouten[1], Gerard J.P. van Breukelen[1], and Martijn P.F. Berger[1]

[1] Department of Methodology and Statistics, University of Maastricht, P.O. Box 616, 6200 MD Maastricht, The Netherlands.

Abstract: Analysis of covariance (ANCOVA) is often used in analyzing data from randomized clinical trials, in which each subject has a pre- and a post-treatment measurement, serving as covariate and outcome, respectively. Also generalized least squares (GLS) related methods, like GEE and multilevel analysis, can be used to analyze such data with pre- and post-treatment measures treated as repeated measures. In this paper, these GLS related methods to estimate a treatment effect, i.e. difference in post-treatment expectation between two groups, are compared with ANCOVA where different assumptions are made about the regression slopes and residual post-treatment variances. We found that ANCOVA is preferred to the GLS method when regression slopes are homogeneous, because of unbiased treatment effect and variance estimators, irrespective whether residual variances are homogeneous or heterogeneous. In case of heterogeneous slopes and homogeneous residual variances, GLS could not be applied with present software and, therefore ANCOVA must be used. Finally, in case of heterogeneous slopes and residual variances, the GLS method is preferred to estimate an *overall* treatment effect, because it takes into account the variability of the pre-treatment mean estimator. But for estimating a *conditional* treatment effect that depends on the pre-treatment value ANCOVA should be used.

Keywords: Analysis of covariance (ANCOVA); Generalized least squares (GLS); Treatment effect estimators.

1 Introduction

Analysis of covariance (ANCOVA) is often applied to data from randomized clinical trials, in which each subject has a pre- and a post-treatment measurement, serving as covariate and outcome, respectively. Also generalized least squares (GLS) related methods, like generalized estimating equations (GEE) and multilevel analysis, can be used to analyze such data.
The purpose of this paper is to compare ANCOVA with these GLS related methods to estimate an unconditional treatment effect, defined as the expected difference in post-treatment values between two treatment groups,

where different assumptions are made about regression slopes and residual post-treatment measures.

2 Methods

Consider a randomized study of G treatments and a quantitative outcome variable. In group g, subject i has a pre-treatment measurement x_{gi} and a post-treatment measurement y_{gi}; $g = 1, 2, \ldots, G$ and $i = 1, 2, \ldots, N_g$. Because of the randomization all groups are assumed to have the same pre-treatment expectation μ_x and variance σ_x^2. The within-group post-treatment expectation μ_g and variance σ_g^2 as well as the covariance between pre- and post-treatment measures $\sigma_{g;\,xy}$ may be influenced by the treatments, indicated by subscript g. The pre- and post-treatment measures are assumed to be bivariate normally distributed with expectation vector $(\mu_x, \mu_g)'$ and covariance matrix V_g. The treatment effect between groups g and g' is defined as the difference in post-treatment expectations, i.e. $\tau_{g,g'} = \mu_g - \mu_{g'}$.
The non-parallel lines ANCOVA model is a conditional model, in which the post-treatment measure y_{gi} is given conditional on the pre-treatment measure x_{gi}, i.e.

$$y_{gi} = \beta_{g0} + \beta_{g1}(x_{gi} - \mu_x) + \epsilon_{gi}, \tag{1}$$

where β_{g0} and β_{g1} are the group-specific intercept and slope of the regression line of group g, respectively. The classical ANCOVA model assumes that the errors ϵ_{gi} are normally and independently distributed with mean zero and homogeneous within-group variance σ_ϵ^2. But, in case of heterogeneous slopes β_{g1}, ANCOVA can account for heterogeneous residual variances across the groups by applying model (1) to each group separately. Since the pre-treatment measurement in model (1) is written as a deviation from the population mean, $(x_{gi} - \mu_x)$, β_{g0} is not only the conditional expectation for a subject with an average pre-treatment value, but also the unconditional expectation μ_g. Thus, the unconditional treatment effect $\tau_{g,g'} = \beta_{g0} - \beta_{g'0}$. The classical parallel lines ANCOVA assumes, additional to homogeneous residual variance, homogeneous slopes across the groups, i.e. $\beta_{11} = \beta_{21} = \ldots = \beta_{G1}$. In this case, μ_x can be omitted from the model without affecting the treatment effect estimators. In case of homogeneous slopes, heterogeneous residual variances imply that weighted least squares (WLS) estimation should be used instead of ordinary least squares (OLS) estimation (see e.g. Diggle, Liang and Zeger, 1994).
Equivalent assumptions to the ANCOVA assumptions of homogeneous or heterogeneous slopes and residual variances can be made in multilevel and GEE models. Therefore, the ANCOVA treatment effect estimators and variances can be compared by those obtained from multilevel analysis or GEE under different assumptions about the ANCOVA regression slopes

and residual variances.

3 Results and Conclusions

The covariance parameters of the repeated measures, σ_x^2, σ_g^2 and $\sigma_{g;\,xy}$ as well as the pre-treatment expectation μ_x are generally unknown and have to be estimated.
In case of homogeneous regression slopes, ANCOVA is then preferred to multilevel analysis or GEE, because ANCOVA leads to unbiased treatment effect estimators and variances, irrespective whether residual variances are homogeneous or heterogeneous. The variances of the GLS treatment effect estimators obtained from multilevel analysis or GEE are biased downwards, because they ignore the variability in the estimators of σ_x^2, σ_g^2 and $\sigma_{g;\,xy}$.
In case of heterogeneous regression slopes, ANCOVA as well as the GLS related methods lead to biased variances of the treatment effect estimators, irrespective whether the residual variances are homogeneous or heterogeneous. ANCOVA ignores the variability in the estimator of μ_x, while the variability in the estimators of σ_x^2, σ_g^2 and $\sigma_{g;\,xy}$ is ignored in the variance of the GLS treatment effect estimators. In case of heterogeneous slopes and homogeneous residual variances, ANCOVA is preferred to GLS, because GLS can not be applied with the present software, like SAS (release 8.02). But, in case of heterogeneous slopes and residual variances, GLS related methods may be used to estimate an unconditional treatment effect instead of ANCOVA, because the variability in the estimator of μ_x can be crucial especially when the regression slopes are quite heterogeneous. More details about the treatment effect estimators and variances as well as the biases in the variances can be found in our paper that has been submitted for publication.

4 Discussion

Kenward and Roger (1997) discussed some simulation studies and proposed to use inflated estimated covariance matrices V_g to adjust for the variability in the estimators of σ_x^2, σ_g^2 and $\sigma_{g;\,xy}$, which was ignored in the variance of the GLS treatment effect estimator. In their paper, the adjustment procedure, which is also implemented in standard software like SAS, was successful in reducing the bias in the variance of the GLS treatment effect estimators. But Kenward and Roger also warned that the proposed adjustment procedure may not be so successful in general as it was in their paper.
The present paper discussed the estimation of unconditional treatment effects, but treatment effects can depend heavily on the pre-treatment value

in case of heterogeneous regression slopes (see e.g. Fleiss, 1986). Such conditional treatment effects can be important for a patient with a particular pre-treatment value. Then, ANCOVA is also preferred to multilevel analysis or GEE in case of heterogeneous regression slopes and residual variances.

References

Diggle, P.J., Liang, K-Y., and Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Clarendon Press.

Fleiss, J.L. (1986). *The Design and Analysis of Clinical Experiments*. New York: Wiley.

Kenward, M.G. and Roger, J.H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, **53**, 983–997.

Sales Forecast for a Pharmaceutical Product Based on Optimal Allocation of Sales Calls and Product Samples

Alex Yaroshinsky

[1] Connetics Corporation, 3290 West Bayshore Road, Palo Alto, California 94303, USA

Abstract: This paper outlines a methodology for optimal allocation of sales calls and product samples among 7200 physicians. Implementation of this methodology led to a 10in the number of prescribing physicians within a month of implementation.

Keywords: Sales; Forecast; Calls; Samples.

1 Methodology

The objective of this paper is to demonstrate a sales forecasting methodology based on the optimal allocation of sales calls and product samples (call and sampling plans). The purpose of these plans was to increase sales and market penetration by assessing optimal quantities and allocation of calls and samples over time among 7200 physicians. We performed a trend analysis on the new and total prescription data for different physician groups structured according to their market potential and market share. From this analysis, optimal number and allocation of sales calls and product samples was determined by building response surfaces" based on the number of prescriptions (TRx) as a function of number of calls and samples per physician for a given period of time. We refer to Figure 1 for a graphical representation of a "response surface".

2 Implementation

The analysis showed that both, prescription volume per month, and trends over time, were substantially different for physician groups depending on the market share of our products and physician-specific prescription volume within given dermatology market segment. For example, growth for physicians with high market share was limited, and promotional activities were only supporting current market share. Respectively, high sales call frequency and samples were not generating additional prescriptions. On the other hand, higher call volume and increased sample allocation were generating significant incremental prescription volume for doctors with low

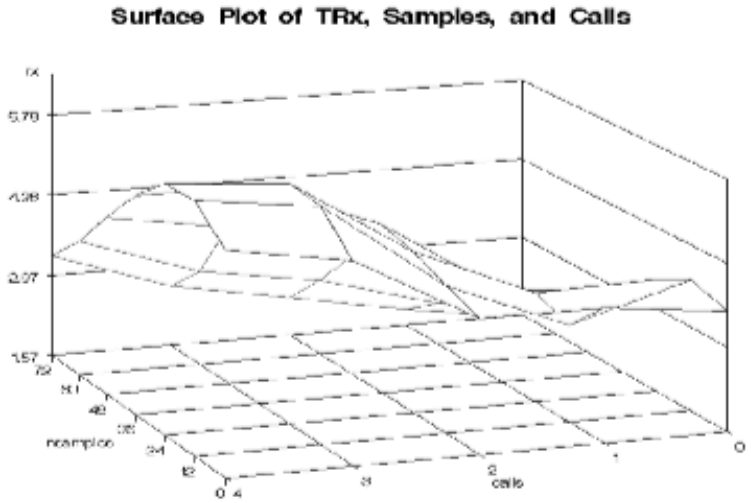


FIGURE 1. *Prescriptions as a function of sales calls and product samples.*

market share, especially those with high potential (determined by the physician's overall prescription volume within particular market segment). Optimal call frequency and sample allocation were implemented by the Connetics'sales force.

Sales forecast was built using dynamic regression model based on the lagged values of dependent and explanatory variables. We refer to Figure 2 for a graphical representation of the 2003 forecast. Prescription volume was used as a response variable with sales call and product sample allocation over time used as explanatory variables. Additionally, overall market volume over time was used to better define seasonality. Current sales force capacity was taken into account: optimal number of sales calls per sales representative could not exceed his/her annual capacity. Budget-driven restrictions were imposed on the optimal sampling program.

3 Results

We were successful in designing optimal sales call and sampling plans as we have experienced significant growth in sales of our products after implementation of this program. Total number of prescriptions per business day in the months following implementation of the plan was growing 10-12a graphical representation of actual sales before and after implementation of a new plan in October. Additionally, as a result of our recommendations,

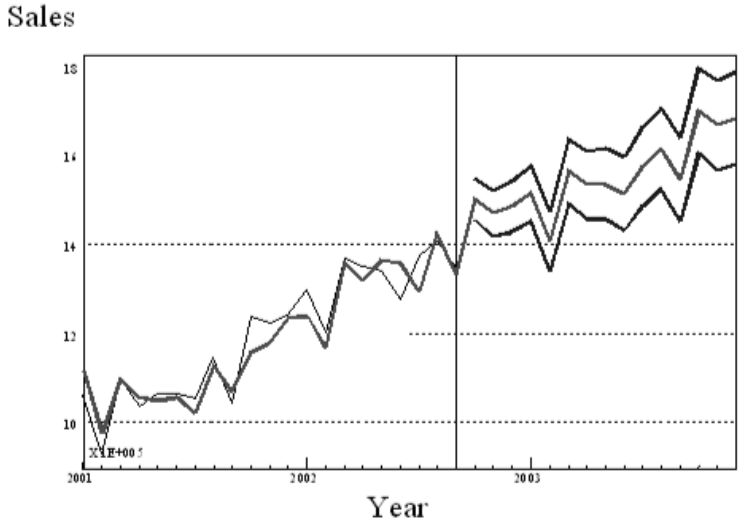


FIGURE 2. *Sales forecast.*

the total number of doctors prescribing our products increased by 18program. These results were further used to generate sales forecast for the first quarter of 2003 and beyond.

Acknowledgments: Special thanks to Mr. Greg Vontz of Connetics Corporation for valuable discussions and recommendations.

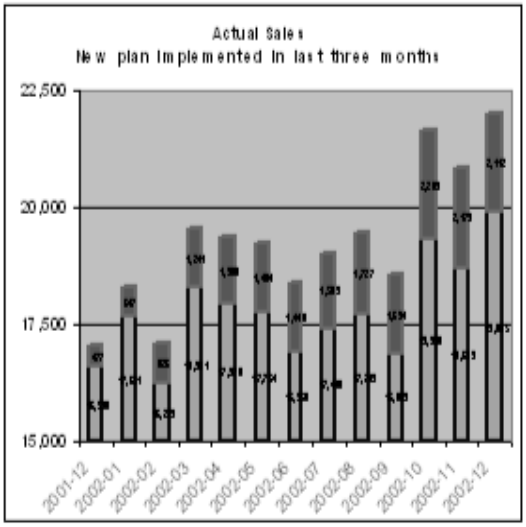


FIGURE 3. *Actual sales before and after implementation of a new plan in October.*

References

Stellwagen, E.A. and Goodrich, R.L. (2000). *Forecast Pro User Manual*. Belmont, MA, USA.

Balkin, S. and Bryden, E. (2002). Non-Linear Forecasting of Physicians' Choice In: *Proceedings of 22nd Int. Symposium on Forecasting ISF2002*. Dublin, Ireland.

Wandering Ideal Point Models For Ranking Data: A Bayesian Approach

Philip L.H. Yu[1] and Jessica H.L. Leung[1]

[1] Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong

Abstract: The wandering ideal point models display ranking data graphically. However, when the number of items to be rank is large, direct maximization of the likelihood function is computationally demanding and numerically unstable. In order to tackle this problem, we adopt a Bayesian approach via MCMC method to estimate the parameters. Simulations are done to demonstrate the proposed estimation method.

Keywords: Wandering ideal point model; Ranking data; Bayesian approach.

1 Introduction

Rankings appear in our everyday life. Marketing research companies want to know the preferences of consumers on different brands of a certain product. Gamblers want to know the ordering of horses in races, but they are not interested in the actual running times of the horses. In addition, rankings eliminates the effects of different scale usage of individual and arbitrarily assignment of the scale in ratings. Proper statistical analysis of these ranking data helps us to study the individual's preference-behaviour.

1.1 Ranking and Ordering

Some people may think that ranking is the same as ordering. Indeed, ordering is a list of items which is according to the acending or decending order-preferences of the judge, while ranking is a list of ranks corresponding to each item. For example, if a judge is presented to $k = 3$ items, tram (item 1), taxi (item 2) and bus (item 3). The ordering $\langle$bus, tram, taxi$\rangle$ means that the judge prefers taxi to tram to bus. We may also code the ordering as $\langle 3,1,2\rangle$. The ranking of the judge is $(R_{tram}, R_{taxi}, R_{bus}) = (R_1, R_2, R_3) = (2, 3, 1)$, because the judge most prefer item 2, item 2 has rank 3. Similarly, the judge least prefer item 3, item 3 has rank 1. Once we know the ordering, we know the ranking, vice versa.

1.2 Interpretation of Ranking Data

Although ranking data is very common, people usually analyze it wrongly. They may treat the discrete ranking data as a continuous scale and then simply use regression or ANOVA to analyze it. Frequently, the very first step of statistical analysis involves data visualization. This helps us to have an insight of the data.

2 Wandering Ideal Point Model

Wandering ideal point model (WIPM), independently developed by De Soete, Carroll and DeSarbo, and Böckenholt and Gaul in 1986, has been widely used in paired comparison of items. For example, if the judge is present to 3 items, he has to make comparisons between items 1 and 2, items 2 and 3, and items 3 and 1. It visualizes individual preferences, so that layman can understand the results easily. However, due to the difficulty in calculating the likelihood function of the wandering ideal point model, no applications on analyzing ranking data using this model for more than 5 items have been found in the literature.

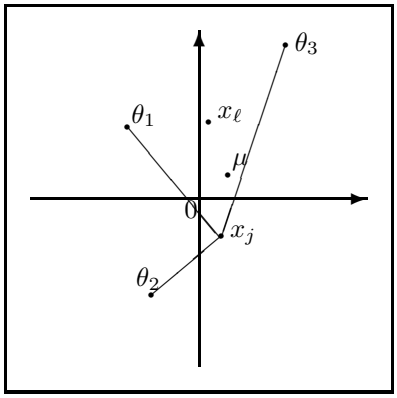


FIGURE 1. An illustration of a 2-dimensional WIPM.

Figure 1 shows a 2-dimensional model. θ_1, θ_2 and θ_3 represent the items to be ranked, while x_j and x_ℓ show the location of judges j and ℓ. The judges sample from the distribution $N_d(\mu, \Gamma)$, where Γ is a diagonal matrix. The distance between x_j and θ_2 is smaller than that of θ_1, which is in turn smaller than that of θ_3. Judge j would prefer item 2 to item 1 to item 3, therefore, ranking of judge j is $(2, 3, 1)$ and ordering $\langle 3,1,2 \rangle$. Similarly, the ranking of judge ℓ is $(3, 1, 2)$. Moreover, a new x is sampled from $N_d(\mu, \Gamma)$ each time when a set of items is presented to the judge. Therefore, the *ideal point* of the judge "*wanders*" around μ from trial to trial, which gives rise to the name of the model.

2.1 Restricitons in Developing the Model

Since the likelihood function is very complicated for the wandering ideal point model, the studies of De Soete *et al.*(1986) and Böckenholt and Gaul (1986) only focused on paired comparisons data. This requires high-dimensional integration and the complexity increases with the number of items to be ranked. The adoption of the classical approach in the parameter estimation restricted the development of the model.

3 Methodology

To overcome the problem discussed in Section 2.1, in our context we adopt a Bayesian approach via Markov chain Monte Carlo methods. Methods proposed here is similar to the one used by Yu and Chan (2001). In order to calculate the posterior distributions, data augmentation (Tanner *et al.*, 1987) is used in our method. Since the ordering of the distances between the items and the judge determines the judge's ranking, we can add the utilities of each judge on the items, which are equal to the negative value of corresponding distances, to the model. Note that we can only observe the ranking, the utilities are unobservable.

3.1 Gibbs Sampler

The Gibbs sampler of the WIPM contains the following 5 steps:

1. Simulate U_{ij} given $x_j,\ \mu,\ \Gamma, \Theta, R_j$
2. Simulate x_j given $U_{ij}, \mu,\ \Gamma, \Theta, R_j$
3. Simulate μ given $U_{ij}, x_j, \Gamma, \Theta, R_j$
4. Simulate Γ given $U_{ij}, x_j, \mu, \Theta, R_j$
5. Simulate Θ given $U_{ij}, x_j, \mu, \Gamma, R_j$

where $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$. The only difference between our approach and that of Yu and Chan (2001) is that the full conditional distribution of Θ is non-standard. Therefore, we proposed to use the Metropolis-Hastings algorithm to generate random variates of Θ. The Metropolis-Hastings algorithm involves a proposed density and a target density, and 3 proposed densities have been tried: (a) Random-Walk (b) Random-Walk with Dependence and (c) Independence. Figure 2 shows the first 20,000 Gibbs iterates of θ_{21} based on these three methods. It can be seen that only Independence Metropolis-Hastings algorithm can get convergence.

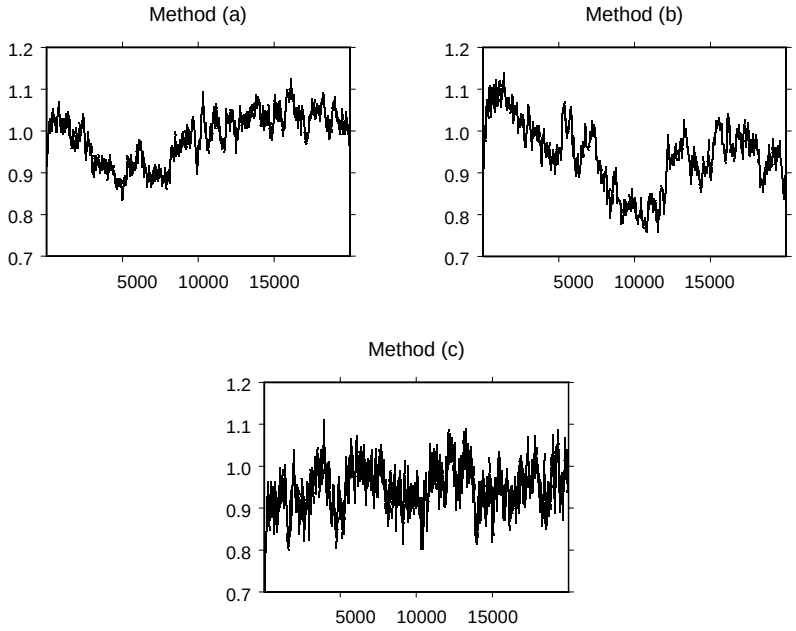


FIGURE 2. θ_{21} against Number of Iterations.

4 Simulation Studies

The true values of the parameters and the results of the simulation studies with 1000 judges and 6 items are shown in Table 1. The number of iterations is 20000 and the burn-in period is 10000 in the Gibbs Sampler. The Independence Metropolis-Hastings Algorithm is chosen for simulating Θ. The initial value for μ is $(1,1)'$, Γ is $\mathbf{I}$ and θ's are $(0,0)'$. Proper but vague conjugate prior distributions were used. The results are satisfactory.

5 Conclusion

The proposed method is efficient in estimating the parameters for the wandering ideal point model for ranking data. The CPU time runs on Compaq Proliant system was about 10 minutes in our simulated studies.

TABLE 1. *Table of Simulation Results.*

Parameter	True Value	Posterior Mean	S.D.	90% Interval	
$\mu_{1,1}$	0.4	0.3998	0.0531	(0.3413,	0.5197)
$\mu_{2,1}$	0.6	0.5913	0.0438	(0.5192,	0.6648)
$\Gamma_{1,1}$	1.2	1.3013	0.1491	(1.1913,	1.6722)
$\Gamma_{2,2}$	1.2	1.2619	0.0878	(1.1182,	1.4111)
$\theta_{1,1}$	-0.8	-0.8137	0.0317	(-0.7659,	-0.8506)
$\theta_{1,2}$	-1.0	-0.9944	0.0382	(-1.0543,	-0.9394)
$\theta_{2,1}$	1.0	1.0170	0.0399	(0.9132,	1.0481)
$\theta_{2,2}$	-1.0	-0.9843	0.0281	(-1.0392,	-0.9461)
$\theta_{3,1}$	0.5	0.5293	0.0330	(0.4706,	0.5808)
$\theta_{3,2}$	0.8	0.7899	0.0333	(0.7196,	0.8319)
$\theta_{4,1}$	-1.0	-1.0354	0.0380	(-1.0674,	-0.9436)
$\theta_{4,2}$	0.5	0.4478	0.0300	(0.4124,	0.5113)
$\theta_{5,1}$	0.3	0.3027	0.0228	(0.2629,	0.3381)
$\theta_{5,2}$	-0.5	-0.5036	0.0281	(-0.5437,	-0.4514)
$\theta_{6,2}$	1.2	1.2445	0.0336	(1.1704,	1.2875)

References

Besag, J., Green, R., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastics systems. *Statistical Science*, **10**, 1–19.

Böckenholt, I. and Gaul, W. (1986). Analysis of choice behaviour via probabilistic ideal point and vector models. *Applied Stochastic Models and Data Analysis*, **2**, 209–226.

De Soete, G., Carroll, J.D., and DeSarbo, W.S. (1986). The wandering ideal point model: A probabilistic multidimensional unfolding model for paired comparisons data. *Journal of Mathematical Psychology*, **30**, 28–41.

Tanner, T. and Wong, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528–549.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion and rejoinder). *Annals of Statistics*, **22**, 1701–1762.

Yu, P.L.H. and Chan, L.K.Y. (2001). Bayesian analysis of wandering vector models for displaying ranking data. *Statistica Sinica*, **11**, 445–461.

Partially-overlapping Covariates in Capture-recapture models

Eugene Zwane[1] and Peter van der Heijden[1]

[1] Department of Methodology and Statistics, Utrecht University, P.O. Box 80.140, 3508 TC Utrecht, The Netherlands. Email: e.zwane@fss.uu.nl

Abstract: Registrations in epidemiological studies suffer from incompleteness, thus a general consensus is to use capture-recapture models. Lately there has been a thrust to incorporate covariates which relate to the capture probabilities in order to improve the estimate of population size. In this presentation we evaluate the usefulness of partially-overlapping covariates and furthermore show how data can be analyzed if the covariates are unobserved for some individuals.

Keywords: EM algorithm; Population size estimation; Capture-recapture; Multinomial logit model; Log-linear model.

1 Introduction

The multiple-record systems estimator is widely used to estimate the size of epidemiological populations using several incomplete registrations (or lists). These registrations usually contain a large number of covariates. In most capture-recapture studies observable heterogeneity is usually taken into account by the use of *fully* overlapping covariates. These covariates can be easily incorporated in log-linear or multinomial logit models. Accounting for observable heterogeneity has been shown to minimize the bias of the estimate of the population size (see Alho, 1990).
In this presentation we evaluate the usefulness of partly overlapping covariates in the multiple system estimator. There is little or no information on the use of non-overlapping covariates. If these covariates are related to the *inclusion* probabilities in-line with the problem with fully overlapping covariates we presume that any estimation (or analysis) excluding them results in a bias.
A simple way to analyze capture-recapture data with non-overlapping covariates will be to replace (impute) each missing value with a single reasonable proxy (alternatives are the mean, random hot-deck, and model-based) for each missing value. This approach is usually called single imputation as only one value us assigned to each missing value. Once the data are imputed, capture-recapture methods incorporating covariates can then be used. The main deficiency of single imputation is that the single value im-

TABLE 1. *Simple problem 2*

	List 2		
List 1	Not included	Included	
Not included	$n_{0(++)}$	$n_{2(+1)}$	$n_{2(+2)}$
Included	$n_{1(1+)}$	$n_{3(11)}$	$n_{3(12)}$
	$n_{1(2+)}$	$n_{3(21)}$	$n_{3(22)}$

puted underestimates the true variability as the imputed value is assumed known with certainty rather than missing in the analysis.
Therefore we take another approach. We will assume that the covariates are missing at random (MAR) in the sense of Little and Rubin (1987); that is, the probability of missingness depends only on the observed data (including the response). This is a reasonable assumption here, because almost all of the missingness is due to unasked questions. Thus we assume that the missingness provides no information about the underlying process, implying that the missing data mechanism is ignorable (see Little and Rubin, 1987). If these assumptions hold, the data can be *efficiently* analyzed using methods for data that are MAR, for example, the EM algorithm and multiple imputation. For this presentation we concentrate on problems where all the covariates are categorical.
In Section 2 we illustrate how the EM algorithm can be implemented for simple cases, that is, where all the covariates are categorical. We show a scenario where ignoring the missing covariates and using traditional methods would result in unbiased estimate of the population size. An analysis of real data on neural tube defects with three overlapping registrations which are incomplete, with both fully overlapping and non-overlapping covariates is analyzed in Section 3. We conclude with a discussion and future work in Section 4.

2 Illustration of EM Algorithm

Assume that we have two *binary* covariates and two lists. Further assume that each list measures one covariate (list 1 measures covariate **A** and list 2 covariate **B**). This scenario is summarized in table 1 (indices for covariates shown in brackets). Note that if all cases with non-overlapping covariates are dropped, the problem is unidentified (or has no solution). If we ignore the covariates the estimate of the numbers missed by all lists is given by;

$$n_{0(++)} = \frac{n_{1(++)} n_{2(++)}}{n_{3(++)}}. \tag{1}$$

If we assume the covariates are related to the inclusion probabilities, the numbers missed using the EM algorithm is,

$$\begin{aligned} n_{0(11)} &= \frac{\left[n_{1(1+)} \times \frac{n_{3(11)}}{n_{3(11)}+n_{3(12)}}\right]\left[n_{2(+1)} \times \frac{n_{3(11)}}{n_{3(11)}+n_{3(21)}}\right]}{n_{3(11)}}; \\ &= \frac{n_{1(1+)}n_{2(+1)}n_{3(11)}}{n_{3(1+)}n_{3(+1)}}; \end{aligned}$$

Similar equations can be made for $n_{0(12)}$, $n_{0(21)}$, and $n_{0(22)}$. These equations hold in the general case, that is when the covariates are dependent. However, if the two covariates are independent, then one could use equation 1 to arrive at an unbiased estimate of the number missed (and population size). This is because under independence,

$$\frac{n_{3(ij)}}{n_{3(i+)}n_{3(+j)}} = \frac{1}{n_{3(++)}}, \qquad i,j = 1,2.$$

This shows that using the covariates (and using the EM algorithm) offers an alternative to the independence model.
Another interesting example is in the three list scenario, where there are two covariates, **A** and **B**: List 1 measures both variables, list 2 measures covariate **A** and list 3 measures covariate **B**. In this example, unlike the previous example, models with the main covariate effects also result in a different estimate of the population size compared to the model excluding covariates. In this case utilizing the partly overlapping covariates provides the researcher with a rich choice of models, which can be discriminated using the Aikake Information Criterion (AIC) or likelihood ratio test.
The variance estimator and confidence interval associated with the estimator of the population size can be constructed using the parametric bootstrap procedure. In most instances the capture counts are approximately distributed as a multinomial distribution. Thus a bootstrap sample is generated from a multinomial distribution with population total equal to the estimated population size and probabilities equal to the fitted probabilities. Note that the cells corresponding to the numbers missed have non-zero probabilities, resulting in the estimator not conditional on the observed sample size.

3 Application

In the Netherlands cases with neural tube defects are registered in several national and regional data bases, and none of these databases include all cases of neural tube defects. For this analysis we use three registrations. The first (R_1) is a registry for low risk pregnancies and birth in primary care, the second (R_2) registers births in secondary care, and the third (R_3)

registers admission and re-admissions of newborns to paediatric department within the first 28 days of life.
In each of the three registries duration of pregnancy (in weeks) and birth (or delivery) weight of the child (in kilograms) is recorded (full overlapping covariates). R_1 and R_2, also have information on age of mother, parity of child, and ethnic group which are not measured in R_3.
We will use data from 1992 to 1998 with 1446 cases. The analysis also has to take into account that children with a pregnancy duration below 24 weeks (abortions) cannot be observed in LNR. We have previously shown how data from registration emanating from different populations can be analyzed using the EM algorithm and the same technique will be used here. The data have 1278 observations with fully observed covariates and 98 observations in LNR only. The other 110 observations have item missing values but none have missing values on all covariates.
To illustrate the EM algorithm in this example the covariates used are birth weight of child (BW: 0 if $<$ 1 kg, 1 if $\geq$ 1 kg), pregnancy duration (PD: 0 if $<$ 24 weeks and 1 if $\geq$ 24 weeks), ethnic group (ETN: 0=Dutch, 1=otherwise), parity (PRT: 0 if $\leq$ 2 children and 1 if $>$ 2 children), and age of mother (MOM: 0 if $<$ 35 years and 1 if $\geq$ 35 years). This analysis will be compared with results from an analysis utilizing only variables observed in all the registrations, that is, pregnancy duration (in weeks) and birth (or delivery) weight of the child (in kilograms).
The models fitted to the data set and corresponding estimates of population size are shown in table 2. In this case the best-fitting model with *fully* overlapping covariates results in a comparable estimate of the populations size to the model including partially overlapping covariates.

4 Conclusions and Discussions

In this article we generalize the multiple systems estimator with categorical covariates to cases where the covariates are not necessarily measured in all registrations. This is accomplished by using the EM algorithm. We show that in the two list case if the covariates are independent in some cases analysis can be performed using the traditional multiple systems estimator. Note that, if there is only one covariate and that covariate is measured by only one registration, it is impossible to measure heterogeneity due to that covariate and it can safely be ignored.
For mixed categorical and continuous covariates we envisage that using multiple imputation (see Schafer, 1997) would be more suitable, but this is a subject of further research.

TABLE 2. *Selected models with deviance and AIC*

	Design matrix	# of par's	AIC	$\hat{N}$
Without covariates				
1	$R_1 + R_2 + R_3 + Year$	10	4491	2423
2	$1 + (R_1 + R_2 + R_3) \times Year$	28	4448	2396
3	2 + H1	29	4436	3292
Including covariates				
4	1 + BW	11	4326	2423
5	4 + PD	12	4151	2277
6	5 + ETN+PRT+MOM	15	1790	2277
7	6 + (BW+PD+ETN+PRT+MOM)$\times Year$	45	1825	2277
8	$6 + (R_1 + R_2 + R_3) \times Year$	33	1746	2250
9	8 + (BW+PD+ETN+PRT+MOM) $\times(R_1 + R_2 + R_3)$	47	1555	2155
10	9 + H1	48	1543	3002
11	10 + H1$\times Year$	54	1552	3072
12	10 + (BW+PD+ETN+PRT+MOM)$\times$H1	52	1549	3046
13	$9 + (R_1 : R_2 + R_1 : R_3 + R_2 : R_3)$	50	1541	3149
14	$13 + (R_1 : R_2 + R_1 : R_3 + R_2 : R_3) \times Year$	68	1548	4474
15	13 +(BW+PD+ETN+PRT+MOM)$\times$ $(R_1 : R_2 + R_1 : R_3 + R_2 : R_3)$	62	1560	2845
16	8 + [PD $\times$ (BW+ETN+PRT+MOM)] $\times$ $(R_1 + R_2 + R_3)+$ $(R_1 : R_2 + R_1 : R_3 + R_2 : R_3)$	62	736	3266
Including fully observed covariates only (best model)				
17	5 +(BW$\times$PD)$\times$ $[(R_1 + R_2 + R_3)+$ $(R_1 : R_2 + R_1 : R_3 + R_2 : R_3)] +$ $(R_1 + R_2 + R_3) \times Year$	41	3090	3165

References

Alho, J. (1990). Logistic regression in capture-recapture models. *Biometrics*, **46**, 495–504.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall.

Author Index