Proceedings of the 21st International Workshop on Statistical Modelling

July 3–7 2006, Galway, Ireland

John Hinde, Jochen Einbeck, John Newell
 editors

ISBN 1-86220-180-3
Editors:

John Hinde
Statistics Group
Department of Mathematics
National University of Ireland, Galway
Galway
Ireland

Jochen Einbeck
Statistics Group
Department of Mathematics
National University of Ireland, Galway
Galway
Ireland

John Newell
Statistics Group
Department of Mathematics
National University of Ireland, Galway
Galway
Ireland

Printed by Corrib Printers in Galway, Ireland
Scientific Programme Committee

- Alan Agresti (Gainesville, USA)
- Jim Booth (Cornell, USA)
- Jochen Einbeck (Galway, Ireland)
- Linda Haines (Cape Town, South Africa)
- John Hinde (Galway, Ireland)
- John Newell (Galway, Ireland)
- Vicente Núñez-Antón (Bilbao, Spain)
- Pedro Puig (Barcelona, Spain)
- Jeff Simonoff (New York, USA)
- Julio Singer (São Paulo, Brazil)
- Ernst Wit (Lancaster, UK)

Local Organising Committee

- Jochen Einbeck (Galway, Ireland)
- Brian Francis (Lancaster, UK)
- John Hinde (Galway, Ireland)
- Emma Holian (Galway, Ireland)
- Gilbert MacKenzie (Limerick, Ireland)
- John Newell (Galway, Ireland)
- Noreen Quinn (Galway, Ireland)
Sponsors:

• Arnold, UK
• Cambridge University Press, UK
• National University of Ireland, Galway, Ireland
• Oxford University Press, UK
Preface

This proceedings volume contains the papers presented at the 21st International Workshop on Statistical Modelling held in Galway, Ireland, 3-7 July 2006. The Workshop aims to bring together statisticians, researchers and all who are interested in the development and application of statistical models, to provide a forum to stimulate new directions and collaborative work.

The Workshop arose out of two GLIM conferences in the U.K. in London (1982) and Lancaster (1985) and a number of short courses in the early 1980’s that sought to publicise, and popularise, the use of generalized linear models. A group of European statisticians saw both the opportunity and the need for a regular meeting that would focus on statistical modelling in an informal workshop environment, aimed specifically at applied statisticians. The inaugural meeting was held in Innsbruck in 1986 and since then the Workshop has moved around Europe with one trip to the USA and last year’s meeting in Australia; Innsbruck, Austria (1986), Perugia, Italy (1987), Vienna, Austria (1988), Trento, Italy (1989), Toulouse, France (1990), Utrecht, Netherlands (1991), Munich, Germany (1992), Leuven, Belgium (1993), Exeter, UK (1994), Innsbruck, Austria (1995), Orvieto, Italy (1996), Biel/Bienne, Switzerland (1997), New Orleans, USA (1998), Graz, Austria (1999), Bilbao, Spain (2000), Odense, Denmark (2001), Chania, Greece (2002), Leuven, Belgium (2003), Florence, Italy (2004), and Sydney, Australia (2005). Future meetings will be in Barcelona (2007) and Utrecht (2008). The meeting has evolved over the last 21 years and has been complemented by other related initiatives. The associated journal Statistical Modelling is now in its 6th year and more recently, in 2003, the Statistical Modelling Society was formed to provide an umbrella organisation for these activities, see the webpage www.statmod.org.

The spirit of the Workshop has always concentrated on papers that are both motivated by real-life data and that also make novel contributions to the subject. Statistical modelling is an important aspect of many diverse disciplines and has a role wherever there is quantitative information to be analysed. The Workshop has provided a forum for the exchange of ideas over many different subject areas and encouraged cross-fertilization. By bringing together scientists from different countries with different backgrounds and different experiences, it has fostered collaboration and spawned new developments. The meeting has also always encouraged the participation of students and those at an early stage of their career, giving opportunities for presentation and time for discussion and interaction between junior and senior scientists. Special emphasis is given to student contributions with awards for the best student presentations.
The scientific programme is characterized by invited lectures, contributed papers and posters, and a pre-workshop short course. One other key feature of the meeting is that there are no parallel sessions – this helps to provide a workshop atmosphere in which discussion and interaction can thrive.

We are pleased to be hosting the Workshop in Ireland for the first time, organised by the Statistics Group from the Department of Mathematics, National University of Ireland, Galway. We are very grateful for generous financial support from the Registrar’s Office at NUI, Galway. We would also like to thank our Scientific Programme Committee for all their work on the selection of the invited speakers and the reviewing of the contributed papers. The invited topics cover many different aspects of modelling and are an impressive set of papers. We would like to thank these speakers, Gerda Claeskens (K.U. Leuven, Belgium), Garrett Fitzmaurice (Harvard School of Public Health, Boston, USA), John Haslett (Trinity College Dublin, Ireland), Leonard Held (LMU, Munich, Germany), and Byron Morgan (University of Kent, UK) for accepting our invitation and preparing such interesting talks. We are also delighted that Ernst Wit (Lancaster University, UK) agreed to present a one-day short course on Statistics for Microarrays: a one day tour based on his book with John McClure, Statistics for Microarrays: Design, Analysis and Inference (Wiley 2004). Finally, our special thanks go to all of the authors who have contributed to the second and main part of this volume, for their care in preparing their manuscripts and giving us a view of the diversity and relevance of statistical modelling in modern research.

We are sure that you will find much to interest and stimulate you in this volume and hope that you have an enjoyable and fruitful stay in Galway.

John Hinde
Jochen Einbeck
John Newell
Galway, May 2006
Contents

Part I: Invited Papers

G. Claeskens: On focussed and less focussed model selection ........ 3
G.M. Fitzmaurice: Methods for handling dropout in longitudinal data analysis ................................................................. 14
J. Haslett et al.: Modelling temporal uncertainty in palaeoclimate reconstructions ......................................................... 26
L. Held: Quantitative assessment of probabilistic forecasts with applications in epidemiology ............................... 38
B.J.T. Morgan: New methods for including covariates in models for the survival of wild animals ....................... 50

Part II: Contributed Papers

M. Amaral Turkman et al.: A Full Bayesian Approach to Physical Maps of Circular Genomes - Application to the Genome of O. oeni strain GM ......................................................... 65
C. Armero et al.: Bayesian evaluation of times to diagnosis in women with mammograms indicating suspicion of malignancy .............................. 70
T.C. Bailey and T.J. McKinley: Spatial survival modelling and the 2001 UK foot-and-mouth disease epidemic .............. 74
S. Barry and A. Bowman: Longitudinal models for shape data 82
S. Cakmak et al.: A measure of progress in air quality management: air health indicator ............................................. 90
M.L. Calle et al.: Analyzing sequential spells via conditional survival models ............................................................. 94
J. Castillo and J. Daoudi: Volatility control on the tails of returns .................................................................................. 102
M. CAZZARO and R. COLOMBI: Marginal modelling of contingency tables with recursive logits and odds ratios ......................... 107
S. CECERE ET AL.: Association structure and ranking of emergence times of permanent teeth: modelling the covariance matrix with interval-censored data ........................................... 111
M.J.L.F. CRUYFF ET AL.: A log-linear randomized-response model to account for cheating ................................................. 119
R. DAHYOT: Bayesian inferences for object detection ............. 127
T. DASU ET AL.: Statistics for network management ............ 131
L.K. DEBUSHO AND L.M. HAINES: Optimal design for linear mixed models ................................................................. 138
M. DURBAN ET AL.: Mixed models, array methods and multidimensional density estimation ................................. 143
P.H.C. EILERS AND B.D. MARX: Multidimensional density smoothing with P-splines .............................................. 151
G. ESCARAMÍS AND J.L. CARRASCO: Approximations for the MSE of random effects in a GLMM estimated by PQL .......... 159
G. ESCARELA ET AL.: Modelling non-Gaussian time series with a mixture copula transition model ................................. 164
J.S. FENLON AND M.J. FADDY: Markov modelling of temporally-ordered functional response data ............................... 172
H.K. FINSELBACH AND A.J. WATKINS: The reliability of censored reliability analyses .................................................... 182
D. FIRTH AND H.L. TURNER: Generalized nonlinear models, and over-parameterized representation .................................. 190
M. GANJI ET AL.: A longitudinal transition ordered response model for analyzing the change of attitudes to gender roles .... 193
S. GREVEN ET AL.: Additive mixed models with P-splines ........ 201
I.D. HA ET AL.: Model selection for frailty structures ............ 208
M.J. HAWKINS ET AL.: An empirical model and scenario analysis of nitrous oxide emissions from a fertilised and grazed temperate grassland site ..................................................... 216
K. HEINER AND J. HINDE: Exponential family mixtures in total quality health care monitoring ........................................ 222
<table>
<thead>
<tr>
<th>Title</th>
<th>Author(s)</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>The zero-adjusted Inverse Gaussian distribution as a model for insurance claims</td>
<td>G. Heller et al.</td>
<td>226</td>
</tr>
<tr>
<td>A multivariate logistic regression model for randomized response data</td>
<td>A. van den Hout et al.</td>
<td>234</td>
</tr>
<tr>
<td>Prediction of risk of sepsis based on deviance in MGLM</td>
<td>Z. Hrdličková</td>
<td>242</td>
</tr>
<tr>
<td>Modelling of recurrence of adenomas: a latent variable approach</td>
<td>C.-H. Hsu</td>
<td>246</td>
</tr>
<tr>
<td>Simultaneous analysis of variables observed in a series of trials</td>
<td>Z. Kaczmarek and E. Adamska</td>
<td>252</td>
</tr>
<tr>
<td>A latent variable model for Australian student’s creativity</td>
<td>H. Kandilorou et al.</td>
<td>257</td>
</tr>
<tr>
<td>Linear mixed-effects modelling of panel data: a cautionary note</td>
<td>I. Kazemi</td>
<td>263</td>
</tr>
<tr>
<td>Survival models for recurrent events in the estimation of the effect of badger culling on tuberculosis in Irish cattle herds</td>
<td>G.E. Kelly and J. Condon</td>
<td>271</td>
</tr>
<tr>
<td>Modelling mortality data on the Lexis diagram</td>
<td>J. Kirkby and I. Currie</td>
<td>279</td>
</tr>
<tr>
<td>A nonparametric multi-state model for the analysis of human sleep</td>
<td>T. Kneib and A. Hennerfeind</td>
<td>286</td>
</tr>
<tr>
<td>Bias reduction and shrinkage in multinomial logit models</td>
<td>I. Kosmidis</td>
<td>294</td>
</tr>
<tr>
<td>Computationally efficient spatially-adaptive penalized splines</td>
<td>T. Krivobokova et al.</td>
<td>303</td>
</tr>
<tr>
<td>A model for combining evidence in several t-statistics to estimate an overall standardized effect</td>
<td>E. Kulinskaya and R.G. Staudte</td>
<td>309</td>
</tr>
<tr>
<td>Bayesian multi-dimensional density estimation with P-splines</td>
<td>P. Lambert and P.H.C. Eilers</td>
<td>313</td>
</tr>
<tr>
<td>Applications of profile confidence intervals for contingency table parameters</td>
<td>J.B. Lang</td>
<td>321</td>
</tr>
<tr>
<td>Product partition models and clustering</td>
<td>V. Livingstone and C. Jordan</td>
<td>330</td>
</tr>
<tr>
<td>Estimating the probability distribution of gamblers by observing betting practices</td>
<td>C. Lynch and D. Barry</td>
<td>336</td>
</tr>
</tbody>
</table>
Y.C. MacNab and P. Gustafson: Spline smoothing in Bayesian disease mapping .......................... 340
J. Maringwa et al.: Analysis of a cardiovascular safety experiment with longitudinal data using penalized smoothing splines 346
J.M. Mendes and K.F. Turkman: Bayesian hierarchical models for hake stocks .......................... 354
A. Mercatanti: A constrained maximum likelihood estimation for relaxing the exclusion restriction in causal inference ........... 359
M.D. Militino et al.: Analysing the performance of alternative unemployment estimators in small areas .................. 363
J. Morales et al.: Bayesian modelling of finite-source queueing systems .................................... 369
D. Morris et al.: Bayesian factor analysis applied to microarray data ........................................ 373
D. Morris and D. Ombredane: Modelling the spatial dynamics of a population of trout using multistate capture-recapture methods 380
V.M.R. Muggeo: Modelling threshold parameter as a function of covariates in segmented regression ............. 389
V. Muntés-Mulero et al.: Modelling genetic methods for data search strategies with ANOVA .................... 393
G. Neubauer and H. Friedl: Modelling sample sizes of frequencies ........................................... 401
M. O'Kelly: A comparison of SAS and R software for performing robust regression .......................... 409
G. Oskrochi and D. Shamley: Analysing the effect of breast cancer treatment on shoulder complexity: a multivariate random effects modelling approach ................................ 414
C. Proust and H. Jacqmin-Gadda: A nonlinear latent class model for joint analysis of multiple longitudinal outcomes and a clinical event .................................................. 422
P. Puig: Appropriate statistical models for analysis of ratios ...... 427
V. Purutçuoğlu and E. Wit: Stochastic simulation of the MAPK signalling pathway .......................... 431
W. Rea et al.: Identification of level shifts in stationary processes .............................................. 438
W.J. Reed: Analyzing forest fire scar data ................................. 442
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>M. Salter-Townshend and J. Haslett: Modelling zero inflation of compositional data</td>
<td>448</td>
</tr>
<tr>
<td>C. Scarrott et al.: Extreme value modelling of reactor risk</td>
<td>457</td>
</tr>
<tr>
<td>M.G. Schimek and T. Pavlík: A comparative study of false discovery rate procedures in R</td>
<td>466</td>
</tr>
<tr>
<td>N. Sofroniou et al.: Analyzing Irish suicide rates with mixture models</td>
<td>474</td>
</tr>
<tr>
<td>J. Tresl and D. Blatná: Statistical analysis of selected European stock indices</td>
<td>482</td>
</tr>
<tr>
<td>G. Tutz and J. Ulbricht: An alternative approach to regularization and variable selection in high dimensional regression modelling</td>
<td>486</td>
</tr>
<tr>
<td>M.D. Ugarte et al.: Restricted predictors for small area totals</td>
<td>494</td>
</tr>
<tr>
<td>F. Vaida and S. Blanchard: Conditional AIC for nonlinear mixed effects models</td>
<td>499</td>
</tr>
<tr>
<td>R. Xu et al.: Using profile likelihood for model selection with application to proportional hazards mixed models</td>
<td>505</td>
</tr>
</tbody>
</table>

Author Index

Author Index
Part I

Invited Papers
On focussed and less focussed model selection

Gerda Claeskens

1 ORSTAT and University Centre for Statistics, K.U.Leuven, Naamsestraat 69, B-3000 Leuven, Belgium

Abstract: Model selection usually provides models without specific concern about for which purpose the selected model will be used afterwards. The focussed information criterion, FIC, is developed to select a best model for inference for a given estimand. For example, in regression models the FIC can be used to select a model for the mean response for each individual subject in the study. This can be used to identify interesting subgroups in the data. Sometimes the FIC is considered too much focussed. We rather would want to select a model that performs well for a whole subgroup, or even for all of the subjects in the study. We explain how to make the focussed information criterion a little less focussed via weighting methods.

Keywords: Focussed information criterion; model selection.

1 Description of the dataset

The dataset considered for discussion in this paper is from the Wisconsin Epidemiologic Study of Diabetic Retinopathy (Klein et al., 1984). It provides information to study diabetic retinopathy as a function of several measurements. The dataset consists of patient information for 343 women and 348 men. The binary outcome variable \( Y = 0 \) indicates whether there is no or only mild nonproliferate retinopathy on both of the eyes. An outcome value \( Y = 1 \) is obtained when there is moderate to severe nonproliferate retinopathy, or proliferate retinopathy for at least one of the eyes. Other variables are: \( x_1 \): the duration of diabetes in years; \( z_1 \): indicator for presence of macular edema in at least one eye; \( z_2 \): the percentage of glycosylated hemoglobin; \( z_3 \): the body mass index; \( z_4 \): pulse rate in beats per 30 seconds; \( z_5 \): sex, using 1 for male and 0 for female; \( z_6 \): indicator for presence of urine protein; and \( z_7 \): area of residence (1=urban, 2=rural).

We fit a logistic regression model to these data. In an earlier analysis of this dataset it is found that duration of diabetes is an important variable (see for example Claeskens, Croux and Van Kerckhoven, 2006), therefore, in this example, we include this variable in all of the models we consider, as well as an intercept term. The model takes the form

\[
\text{logit}P(Y = 1) = \beta_0 + \beta_1 x_1 + \sum_{s \in S} \beta_s z_s,
\]
where \( S \) is an index set, containing the variables contained in the corresponding model. For example, the biggest model considered has \( S = \{1, 2, \ldots, 7\} \), containing all of the variables \( z_1, \ldots, z_7 \). The empty set \( S = \emptyset \) corresponds to fitting the smallest model \( \logit P(Y = 1) = \beta_0 + \beta_1 x_1 \). The model selection questions we pose are which of the variables \( z_j \) are to be included in the model, thereby distinguishing between an individual model search, and a search for the subset of women separately.

2 Individual model searches by FIC

In this section we explain how to obtain a very focussed model selection, subject specific. This follows the approach as in Claeskens and Hjort (2003). We start with defining the focus parameter. For ease of explanation we use the logistic regression model as in the example in the previous section. As a focus parameter we take the linear predictor

\[
\mu = x^t \beta + z^t \gamma = \logit \{ E(Y|x,z) \}.
\]

It is important to note that this focus parameter changes for each different set of covariate values \((x,z)\). We define the design matrices \( X \) and \( Z \), of dimension \( n \times p \) and \( n \times q \) respectively. In the above example we have \( p = 1 \) and \( q = 7 \). The \( i \)th row of \( X \) consists of \((1, x_{1i}, \ldots, x_{p-1,i})\), and the \( i \)th row of \( Z \) consists of \((z_{1i}, \ldots, z_{qi})\), for \( i = 1, \ldots, n \). Here we assume that the intercept is contained in all models, of course this can be changed. The Fisher information matrix in the largest model is crucial for the FIC, and in particular the lower right submatrix of the inverse Fisher information matrix, which we denote by \( K_n \). For logistic regression

\[
K_n = n (Z^t V (I - X(X^t V X)^{-1} X^t V X)Z)^{-1},
\]

where \( V = \text{diag}\{p_i(1-p_i)\} \), with \( p_i = P(Y_i = 1|x_i, z_i) \) according to the logit model including all variables. For practical use, we insert estimators for the unknown parameters, obtained from the full model. A third ingredient is a vector

\[
\omega = Z^t V X (X^t V X)^{-1} x - z,
\]

also clearly depending on the covariate values \((x,z)\). This, together with estimators \( \hat{\gamma} \) obtained in the largest model, is all that is needed to obtain the value of the focussed information criterion, for the purpose of model selection at the exact covariate position \((x,z)\). The value of FIC is obtained for each model indexed by a set \( S \) separately. Each time we select the corresponding rows and columns of the matrix \( K_n^{-1} \). For example, if \( S = \{2, 5\} \), we select from \( K_n^{-1} \) the \( 2 \times 2 \) submatrix containing entries at the 2nd and 5th rows and 2nd and 5th columns. Algebraically this is denoted by means of a projection matrix \( \pi_S \), which is of dimension \(|S| \times q \) and selects from a vector \( v \) only those components \( v_j \) for which \( j \in S \), which we denote
TABLE 1. Percentage of times that a variable has been selected by the individual model searches.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td>z₁: Edema</td>
<td>0.621</td>
<td>0.693</td>
</tr>
<tr>
<td>z₆: Urine protein</td>
<td>0.446</td>
<td>0.382</td>
</tr>
<tr>
<td>z₄: Pulse</td>
<td>0.201</td>
<td>0.216</td>
</tr>
<tr>
<td>z₂: Hemoglobin</td>
<td>0.149</td>
<td>0.158</td>
</tr>
<tr>
<td>z₃: Body mass index</td>
<td>0.131</td>
<td>0.118</td>
</tr>
<tr>
<td>z₇: Residence</td>
<td>0.070</td>
<td>0.083</td>
</tr>
<tr>
<td>z₅: Sex</td>
<td>0.035</td>
<td>0.034</td>
</tr>
</tbody>
</table>

by $v_S = \pi_S v$. We also used the notation $|S|$ for the number of components in $S$. With this we define $K_{n,S} = (\pi_S K_n^{-1} \pi_S^{-1})^{-1}$, and $G_{n,S} = \pi_S K_n S \pi S K_n^{-1}$, and finally

$$\text{FIC}(S; x, z) = n\omega(\pi_{n,S} g'g(I_q - G_{n,S})I_q - G_{n,S})\omega + 2
\omega^T K_n S \pi S \omega.$$  

The FIC decomposes into a squared bias estimator and a variance estimator (times 2). In cases where the estimated squared bias component happens to be negative, we replace it by zero. Let $S_1, S_2, \ldots$ be the set of index sets for the model search. We obtain a list of corresponding $\text{FIC}(S_1; x, z), \text{FIC}(S_2; x, z), \ldots$. The best model is that one with the smallest value of FIC.

3 Individual model selection for the dataset

The WESDR dataset is used for an individual model search. We perform the model search as described above, where the covariate positions chosen correspond to the subset of 343 women in the study. This means, for each $(x_i, z_i)$ where $i = 1, \ldots, 343$ we compute all $2^7 = 128$ models in an all subsets model search. The best model according to the FIC, is that model for which the corresponding FIC value is the smallest of the 128 FIC values computed for this person. This dataset is quite rich in the sense that it here is not obvious that one model is best for all different subjects. In fact, the picture in Figure 1 shows that 39 models were selected at least once. Three model stand out: they got selected 95, 46 and 35 times respectively. The order of the model numbers is irrelevant. These models include only variable $z_1$, no extra variables, and only variable $z_6$ respectively.

A study of the 39 selected models reveals that $z_1$, presence of macular edema was chosen in 62.1% of the 343 subject specific models, $z_3$ pulse was chosen 20.1%, hemoglobin level $z_6$ came third with 14.9%. Table 1 gives the full results for all seven variables, for the subsets of men and women separately.
FIC model selection for female patients

FIGURE 1. For each subject in the study a separate model search has been performed amongst $2^7 = 128$ different models. The graph shows the model number (from 1 to 128) and indicates for which subjects that particular model was selected. The numbers on top are the most frequent model counts. 39 models got selected at least once.

Let us now take a closer look at the most selected individual model, the model with extra variable $z_1$. Figure 2 summarises the search result for this model across all female individuals in the study. We already know that the model received 95 times rank 1. A closer look reveals that it received 25 times rank 2, 23 times rank 3 and for 230 out of the 343 subjects, corresponding to 67%, it received a rank number no higher than 10 (out of 128).

Figures 1 and 2 clearly illustrate that, at least for some datasets, it may be advantageous to perform a subjectwise model selection. There is no obvious overall winner.
FIGURE 2. Across all 343 female patients in the study, we count the rank number of the most often individually selected model according to a subjectwise FIC, including extra variable $z_1$, amongst all possible subset models. A frequency histogram is shown here. This model received 95 times rank 1, 25 times rank 2, etc.

4 Weighted FIC

An individual model search is sometimes not what one wants, rather, we wish to select a model that is focussed towards estimation of a certain focus parameter, but at the same time the model should perform well for all of the subjects. One strategy is to choose average values for the covariates and go through the model selection step for this hypothetical average subject. This would follow exactly the same steps as an individual model search, and needs no further explanation. In this section we discuss how to construct a weighted version of the FIC. We start with a weighted average quadratic
loss function on the scale of the linear predictor

\[
\sum_{i=1}^{n} w(x_i, z_i) \{ \hat{\mu}_S(x_i, z_i) - \mu_{\text{true}}(x_i, z_i) \}^2.
\]

The weights \( w(x_i, z_i) \) are user-specified. The expected value of this loss function can again be decomposed into a weighted squared bias and weighted variance term. The weighted focussed information criterion \( w\text{FIC} \) is an estimator hereof. To state the exact definition, denote \( W = \text{diag}\{w(x_i, z_i)\} \), and let

\[
\Omega_n = \frac{1}{n} \begin{pmatrix} X'WX & X'WZ \\ Z'WX & Z'WZ \end{pmatrix}, \quad J_n = \frac{1}{n} \begin{pmatrix} X'VX & X'VZ \\ Z'VX & Z'VZ \end{pmatrix}.
\]

We denote by \( J_{n,S} \) the corresponding submatrix for the model indexed by \( S \). The matrix \( \Omega_n \) only differs from the Fisher information matrix \( J_n \) by the use of the weight matrix \( W \) instead of the logistic weights \( V \). Further, define for each index set \( S \)

\[
F_{n,S} = \begin{pmatrix} (X'VX)^{-1}X'VZ(I_q - G_{n,S}) \\ -(I_q - G_{n,S}) \end{pmatrix}
\]

and an extended projection matrix of dimension \((p + |S|) \times (p + q)\),

\[
\tilde{\pi}_S = \begin{pmatrix} I_p & 0_{p,q} \\ 0_{|S|,p} & \pi_S \end{pmatrix}.
\]

The weighted FIC is now computed as follows:

\[
w\text{FIC} = \text{trace}(\Omega_n \tilde{\pi}^t_S (J^{-1}_{n,S} \tilde{\pi}_S)) + \max\{\text{trace}(\Omega_n F_{n,S}(n\hat{\gamma}^t - \hat{\gamma}^{(0)}_n)F_{n,S}^t)\}, 0\}.
\]

The construction with the truncation by zero avoids obtaining a negative bias-squared estimate. For more details and a justification of this derivation we refer to Claeskens and Hjort (2006).

5 Logistic variance weights and a connection to AIC

The weights in the weighted FIC are user defined. One particular type of weights leads not only to a simplification of the \( w\text{FIC} \) formula, but also to a connection to Akaike’s (1974) information criterion. Let us take \( w(x_i, z_i) = p_i(1 - p_i) \), which implies that \( W = V \), and hence that \( \Omega_n = J_n \) and \( \text{trace}(\Omega_n \tilde{\pi}^t_S J^{-1}_{n,S} \tilde{\pi}_S) = p + |S| \), the number of parameters in the model indexed by \( S \). For a positive estimated squared bias, we then get that, for this particular choice of weights,

\[
w\text{FIC} = n\hat{\gamma}^t(K_n^{-1} - K_n^{-1}n\pi_SK_{n,S}\pi_S K_n^{-1})\hat{\gamma} + 2|S| + p - q.
\]
Let us compare this to Akaike’s information criterion

$$\text{AIC}(S) = -2 \sum_{i=1}^{n} \log f(y_i, x_i, z_i, \hat{\beta}, \hat{\gamma}) + 2(p + |S|).$$

The AIC values are obtained for each submodel $S$, including $S = \emptyset$, corresponding to not including any extra variables $z_j$. When subtracting the smallest model’s AIC value from $\text{AIC}(S)$, and performing a one-step Taylor expansion, we find that, for $n \to \infty$,

$$\text{AIC}(S) - \text{AIC}(\emptyset) \xrightarrow{d} -D'K^{-1}_{n}n_{S,K}n_{S,K}^{-1}D + 2|S|.$$

See Claeskens and Hjort (2003, eq. (2.5)). The variable $D$ in the expression above is in practice estimated via $\sqrt{n}\hat{\gamma}$. This immediately shows the connection to the weighted FIC, with weights equal to the logistic variances. Of course, the weighted FIC is able to incorporate other types of weights as well. In the next section, we illustrate this by example on the WESDR dataset, where we choose equal weighting for all female subjects in the dataset. An example of robust downweighting is shown on Hoffstedt’s highway data in Section 7.

6 Weighted FIC model selection for the dataset

To obtain an overall best model according to FIC for the subset of women in the study, we compute for each subset $S$, $w\text{FIC}(S)$ with weight vector $(1/nF) I(\text{female})$ where the weights are indicator variables for women and $nF$ denotes the number of women in the dataset. It is important to note that the values of $w\text{FIC}(S)$ are computed using the complete dataset, we are not splitting the dataset for model selection. Figure 3 shows the 20 smallest (best) $w\text{FIC}$ values. There is not much difference between the first two values, which correspond to the models: (i) including variables $z_1$, $z_4$ and $z_6$ for the best model, and (ii) including variables $z_1$, $z_4$, $z_6$ and $z_7$ for the second best model. The best model corresponds to model number 54 in Figure 1, which was preferred 13 times as individually best model. This particular model is also the one chosen via a –not focussed– AIC model search.

7 Downweighting outlying observations, an example

As an illustration we use Hoffstedt’s highway data, included in R’s library(alr3), see also Weisberg (2005, Section 7.2). This dataset is used to explain the 1973 accident rate per million vehicle miles, as a function of several variables. There are 39 observations. In every model we include an intercept term and $x_1$, the length of the highway segment in miles. Other
variables are $z_1$: average daily traffic count in thousands, $z_2$: truck volume as a percent of the total volume, $z_3$: total number of lanes of traffic, $z_4$: number of access points per mile, $z_5$: number of signalised interchanges per mile, $z_6$: number of freeway-type interchanges per mile, $z_7$: speed limit in 1973, $z_8$: lane width, in feet, $z_9$: width of the outer shoulder on the roadway (in feet), and finally $z_{10}$: an indicator of the type of roadway or the source of funding for the road.

We first fit the full model using the method of M-estimation using Huber’s psi function. This gives a set of residuals. Figure 4 shows the plot of these residuals against the fitted values. We identify five points where the absolute value of the residual is larger than 1.345. This particular cut-off value is also used by Ronchetti and Staudte (1994). We now define

$$w(x_i, z_i) = \begin{cases} 1 & \text{if } |\text{residual}_i| \leq 1.345 \\ 1.345/|\text{residual}_i| & \text{if } |\text{residual}_i| > 1.345, \end{cases}$$

FIGURE 3. The 20 smallest wFIC values for the WESDR dataset where the weights are indicator variables for the subset of women. This is global model search.
and use these as weights to perform a global model search. The five identified residuals, outside the (-1.345, 1.345) range in Figure 4 lead to weights 0.903, 0.568, 0.436, 0.577, and 0.811. All other observations get weight equal to one. This approach is effectively downweighting these 5 influential observations. 

Figure 5 shows the 50 smallest $wFIC$ values. The best value corresponds to the model including $x_1$ and $z_4$. 

Based on robust $C_p$ model selection, Ronchetti and Staudte (1994) support the model which includes, in addition to $x_1$, the variables $z_5$, $z_6$, $z_7$ and $z_{10}$, and also the model with additional variables $z_2$, $z_3$, $z_4$ and $z_9$. For this example, the model chosen by $wFIC$ is more parsimonious.
8 Remarks

It is worth investigating whether an individual model search is useful in identifying “outlying” observations. Outlying can here be understood in the sense of pointing to quite different models than the majority of the other observations.

When outlying observations are identified, they can be downweighted in a weighted model search. Also this approach needs some more investigation.

Claeskens, Croux and Van Kerckhoven (2006) construct a different type of (subjectwise) focussed information criterion. Instead of selecting a model with the goal of minimising the MSE, other FIC expressions are constructed which minimise the prediction error. This strategy is useful to direct model selection even more to a specific task.
References


Methods for handling dropout in longitudinal data analysis

Garrett M. Fitzmaurice

1 Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, U.S.A

Abstract: In longitudinal studies, missing data are the rule not the exception. We briefly review the statistical literature on approaches for handling dropout in longitudinal studies. A fundamentally difficult problem arises when the probability of dropout is thought to be related to the specific value that in principle should have been obtained; this is often referred to as “informative” or “non-ignorable” (NI) dropout. Joint models for the longitudinal outcomes and the dropout times have been proposed to make corrections for non-ignorable dropout. Two broad classes of joint models are reviewed: selection models and pattern mixture models. Selection and pattern mixture modeling approaches each have their own distinct advantages and disadvantages. We describe an alternative approach, based on a "hybrid model", that attempts to capitalize on some of the advantageous features of both modeling approaches. The proposed method is illustrated using data from a clinical trial of contracepting women.

Keywords: Dropout; longitudinal data; missing data.

1 Introduction

Missing data are a common feature of longitudinal designs that complicate their analyses. Subjects may miss scheduled measurement occasions, producing intermittently missing responses, or may become lost to follow-up, resulting in the monotone pattern of dropout. The focus of this paper is on the problem of dropout in longitudinal studies. For example, a year-long clinical trial comparing two doses of an injectable contraceptive regimen (Machin, et al., 1988) randomized 1,151 women to receive either 100mg or 150mg of depot-medroxyprogesterone acetate (DMPA) at the start of each of four 90-day intervals. The women maintained menstrual diaries within each interval, and a binary indicator of menstrual irregularity (amenorrhea) was the outcome of interest for each interval. More than one-third of the women dropped out before completion of the study, and for reasons thought to be related to the outcome of interest. In this paper, we focus on methods for analyzing longitudinal data subject to such non-ignorable (NI) dropout, where dropout is assumed to depend on unobserved responses.
2 Models for Longitudinal Data with Non-Ignorable Dropout

Models for longitudinal data subject to NI dropout can be broadly classified into two categories: selection models and pattern mixture models (Little, 1993). Before reviewing these two approaches, we introduce some notation.

2.1 Notation

For ease of exposition, assume that \( N \) individuals are to be observed at the same set of \( n \) occasions, \( \{t_1, t_2, \ldots, t_n\} \). Let \( Y_{ij} \) denote the response for the \( i \)-th individual at the \( j \)-th occasion; \( Y_{ij} \) can be discrete, ordinal, continuous or count data. Let \( Y_c^i \) denote the \( n \times 1 \) complete response vector, \( Y_c^i = (Y_{i1}, \ldots, Y_{in})' \). In addition, associated with each \( Y_{ij} \) is a \( p \times 1 \) vector of covariates, \( X_{ij} \); these can be grouped into a \( n \times p \) matrix of covariates denoted by \( X_i \). The analytic goal is to make inferences about a set of regression parameters, \( \beta \), in the following model for the mean response,

\[
g(E(Y_c^i|X_i)) = X_i \beta,
\]

where \( g(\cdot) \) is a known link function.

Each individual also has a discrete event time \( D_i \), denoting non-ignorable dropout. Let \( D_i \in \{t_1, \ldots, t_n\} \) denote the last observed measurement occasion and let \( \phi_{ij} = \Pr(D_i = t_j) \). If \( D_i \neq t_n \), the \( i \)-th subject is a “dropout”; otherwise, a study “completer”. Recall that dropout is “non-ignorable” when \( D_i \) depends on unobserved \( Y_{ij} \). Finally, let \( Y_i \) denote the \( n_i \times 1 \) vector of the responses observed on the \( i \)-th individual, i.e., the observed portion of \( Y_c^i \). Thus, the observed data for each subject consist of \( (Y_i^c, D_i, X_i) \); in a longitudinal study, the covariates in \( X_i \) will generally include treatment or exposure group, in addition to time \((t_j)\).

2.2 Selection Models

To correct for potential bias when dropout is non-ignorable, joint models for the distribution of \( Y_c^i \) and \( D_i \) have been proposed. In selection models the joint distribution of \( Y_c^i \) and \( D_i \) is expressed as follows,

\[
f(Y_c^i, D_i|X_i) = f_Y(Y_c^i|X_i) f_{D|Y_c^i}(D_i|Y_c^i, X_i).
\]

Recall that in longitudinal studies, the primary focus is on inferences about \( f_Y(Y_c^i|X_i) \). In contrast, \( f_{D|Y_c^i}(D_i|Y_c^i, X_i) \) plays the role of “nuisance parameters”, which can be ignored only if \( f(D_i|Y_c^i, X_i) \) does not depend upon any missing \( Y_{ij} \)’s (or random effects).

There is an extensive literature on selection models for dropout. For example, Wu and Carroll (1988) proposed a selection modeling approach used...
by many subsequent authors. It assumes that the continuous longitudinal responses follow a simple linear random effects model (e.g., Laird and Ware, 1982) and that the dropout process depends upon an individual’s random slope. Models where the dropout probabilities depend indirectly upon the unobserved responses via random effects are sometimes referred to as shared parameter models. An alternative selection modeling approach was proposed by Diggle and Kenward (1994), who allowed the probability of nonresponse to depend directly on the unobserved outcome rather than on random effects. Selection models, where the nonresponse probabilities depend directly or indirectly (via dependence on random effects) on unobserved responses, have also been extended to discrete longitudinal data (e.g., Baker, 1995; Fitzmaurice, Molenberghs & Lipsitz, 1995; Fitzmaurice, Laird & Zahner, 1996; Molenberghs, Kenward & Lesaffre, 1997; Fitzmaurice, Heath & Clifford, 1996; Ten Have et al, 1998, 2000). Semiparametric selection model formulations, based on inverse probability weighted generalized estimating equations (GEE), have been proposed by Robins, Rotnitzky & Zhao (1995) and Scharfstein, Rotnitzky & Robins (1999).

2.3 Pattern Mixture Models

In pattern mixture models the joint distribution of $Y_{ci}$ and $D_i$ is expressed as follows,

$$f(Y_{ci}, D_i|X_i) = f_D(D_i|X_i)f_{Y|D}(Y_{ci}|D_i, X_i).$$

Recall that in longitudinal studies inferences about $f_{Y|D}(Y_{ci}|D_i, X_i)$ are not usually of main interest. Rather, the primary interest is on inferences about $f_Y(Y_{ci}|X_i)$, obtained by integrating over the distribution of $D_i$.

The literature on pattern mixture models is less extensive. For example, Wu and Bailey (1988, 1989) suggested a pattern mixture modeling approach based on method-of-moments type fitting of a linear model to least squares slopes, conditional on dropout time, and then averaging over the distribution of dropout time. Hogan and Laird (1997) extended this pattern mixture model by permitting censored dropout times, as might arise when there are late entrants to a trial and interim analyses are performed. Follmann and Wu (1995) generalized the Wu and Bailey conditional linear model to permit generalized linear models without any parametric assumption on the random effects. Fitzmaurice and Laird (2000) proposed an extension of the conditional linear model of Wu and Bailey (1988, 1989) and Hogan and Laird (1997) that permits generalized linear models for the conditional mean without conditioning upon any random effects. Other related work on pattern mixture models is described in Rubin (1977), Glynn, Laird and Rubin (1986), Mori, Woodworth and Woolson (1992), Heder and Gibbons (1997), Eksolm and Skinker (1998), Molenberghs et al (1998, 1999), Park and Lee (1999), and Michiels, Molenberghs and Lipsitz (1999). Finally, there is an additional avenue of research on pattern mixture models
that can be distinguished. Little (1993; 1994) has considered pattern mixture models which stratify the incomplete data by the pattern of missing values and formulate distinct models within each stratum. In these models additional assumptions about the missing data mechanism have to be made in order to yield supplemental restrictions that identify the models.

2.4 Relative Merits of Pattern Mixture and Selection Models

Pattern mixture and selection models each have their own distinct advantages and disadvantages. A chief advantage of selection models is that they includes the regression parameters of primary interest in a longitudinal analyses. Also, within selection models, it is straightforward to formulate hypotheses about the dropout process. While assumptions about the dropout process are transparent in selection models, what is less clear is how these translate into assumptions about the distributions of the unobserved outcomes. Furthermore, with selection models identification comes from postulating unverifiable models for the dependence of the dropout process on the unobserved outcomes. However, except in very simple cases, it is difficult to determine the identifying restrictions that must be placed on the model (Glonek, 1998). Finally, selection models are well-known to be sensitive to model assumptions and are computationally demanding.

In contrast, pattern mixture models are often as easy to fit as standard models that assume dropout is ignorable. With pattern mixture models it is immediately clear that the distribution of the outcomes given patterns of dropout is not completely identifiable, since for all but the “completely observed” pattern, certain variables are not observed. Identification comes from postulating unverifiable links among the distributions of the outcomes conditional on the patterns of dropout. But, in contrast to selection models, it is relatively straightforward to determine the identifying restrictions that must be imposed. However, pattern mixture models do have one very important drawback that has, so far, limited their usefulness in many areas of application. The main drawback of pattern mixture models is that the natural parameters of interest are not immediately available; they require marginalization of the distribution of outcomes over dropout patterns. Although the marginal distribution can be obtained by averaging over the distribution of the dropout patterns (e.g., Wu and Bailey, 1989; Hogan and Laird, 1997; Fitzmaurice and Laird, 2000), note that the assumed form for the regression model for the conditional means (e.g., logistic or log-linear) no longer holds for the marginal means when a nonlinear link function has been adopted. As a result, with pattern mixture models it is not possible to parsimoniously describe the effects of covariates on the marginal means in terms of regression coefficients.
3 Marginally-Specified Pattern Mixture Models

To capitalize on some of the advantageous features of both selection and pattern mixture models, we propose an alternative parameterization for pattern mixture models in which the marginal means of the outcomes, rather than the conditional means given dropout patterns, are regressed on the covariates. This circumvents the chief drawback of the standard pattern mixture modeling approach, while still retaining many of the advantages of pattern mixture models that have been highlighted earlier.

Recall that the basic idea underlying the standard pattern mixture model,

\[ f(Y^c_i, D_i|X_i) = f_D(D_i|X_i)f_Y(Y^c_i|D_i, X_i), \]

is stratification by different patterns of dropout. In particular, the distribution of \( Y^c_i \) (given \( X_i \)) for those who dropout must be related to the distribution of \( Y^c_i \) for those who complete the study. Consider the following example where the model for \( Y_{ij} \), conditional on the time of dropout, has the following general form:

\[ g\left[ E(Y_{ij}|X_{ij}, D_i) \right] = Z_{ij}'\beta^* \]

where \( g(\cdot) \) is a known link function (e.g., log or logit), and the design vector \( Z_{ij} \) depends on dropout time, \( D_i \), and also incorporates the covariates \( X_{ij} \). Note that the model for the conditional mean of \( Y_{ij} \) will not be identified unless some (unverifiable) assumptions are made. Recognizing that, without additional information, identification is driven by unverifiable assumptions, these models should be used within the context of a sensitivity analysis. That is, results should always be reported for analyses conducted under a range of plausible assumptions about non-ignorable dropout.

In many longitudinal studies, the parameter of primary interest is not \( \beta^* \). Rather, the target of inference is the marginal expectation of the repeated outcomes, averaged over the distribution of the dropout times,

\[ E(Y_{ij}|X_{ij}) = \mu_{ij} = \sum_{l=1}^{n} \phi_{il} g^{-1}(Z_{ij}^l\beta^*), \]

where \( Z_{ij} \) depends on the dropout patterns, and \( \phi_{il} \) depends on \( X_i \) (or a subset of \( X_i \)). Herein lies the problem with the standard pattern mixture modelling approach: for a non-linear link function, \( g(\cdot) \), if \( g[E(Y_{ij}|X_{ij}, D_i)] = Z_{ij}^l\beta^* \), then \( g[E(Y_{ij}|X_{ij})] \neq X_{ij}'\beta \). For example, if

\[ \logit[E(Y_{ij}|X_{ij}, D_i)] = Z_{ij}^l\beta^*, \]

then

\[ \logit[E(Y_{ij}|X_{ij})] = \log \left( \frac{\mu_{ij}}{1-\mu_{ij}} \right) \]

\[ = \log \left( \frac{\sum_{l=1}^{n} \phi_{il} \exp(Z_{ij}^l\beta^*)}{1-\sum_{l=1}^{n} \phi_{il} \exp(Z_{ij}^l\beta^*)} \right) 
\neq X_{ij}'\beta, \]
and a logistic regression model no longer holds for the resulting marginal means. Thus, the chief problem with the standard pattern mixture model is that the assumed form of the regression model for the conditional means (e.g., logistic or log-linear) no longer holds for the marginal distribution after integration over the distribution of the dropout times. To circumvent this problem we propose marginally-specified models that involve three main components:

(i) Marginal model for mean of $Y_{ij}$: 
\[ E(Y_{ij}|X_{ij}) \]

(ii) Marginal model for dropout pattern, $D_i$: 
\[ f_D(D_i|X_i) \]

(iii) Conditional model for mean of $Y_{ij}$ given $D_i$: 
\[ E(Y_{ij}|D_i, X_{ij}) \]

In component (i) we specify a marginal model for the mean of $Y_{ij}$,
\[ g[E(Y_{ij}|X_{ij})] = X_{ij}'\beta. \]

Component (ii) specifies a marginal model for dropout, $D_i$, that conditions only on the covariates, $X_i$. Depending upon the study design, the multinomial probabilities for dropout, $\phi_i = (\phi_{i1}, ..., \phi_{in})'$, can often be estimated nonparametrically as the sample proportion with each dropout pattern (stratified by exposure or treatment group and, perhaps, by other relevant covariates). Alternatively, we can consider parametric models, e.g., multinomial regression models, for $\phi_i$. Finally, component (iii) specifies a model for the conditional mean of $Y_{ij}$ given $D_i$,
\[ g[E(Y_{ij}|X_{ij}, D_i)] = \Delta_{ij} + Z_{ij}'\beta^*, \]
where $Z_{ij}$ depends on $D_i$ and also incorporates the covariates $X_{ij}$. Note the inclusion of $\Delta_{ij}$ in the linear predictor; $\Delta_{ij}$ is defined implicitly as a function of $\beta, \beta^*, \phi_i$, via the following equation
\[ E(Y_{ij}|X_{ij}) = \mu_{ij} = \sum_{l=1}^{n} \phi_{il} g^{-1}(\Delta_{ij} + Z_{ij}'\beta^*). \]

Finally, we remark that these three components do not fully specify the joint distribution of $Y_{ij}$ and $D_i$; instead, they specify a semi-parametric model.

3.1 Estimation

Once identifying constraints are adopted, $\beta$ (and $\beta^*$) can be estimated via the solution to a set of generalized estimating equations,
\[ \sum_{i=1}^{N} G_i [V_i^{-1} [Y_i - E(Y_i|X_i, D_i)] = 0, \]
20 Methods for handling dropout

where

\[ G_i = \frac{\partial E(Y_i \mid X_i, D_i)}{\partial \theta}, \text{ and } \theta = (\beta', \beta^*) , \]

and \( V_i \) is an appropriate weight matrix. The solution to the GEE also requires solving for implicitly defined \( \Delta_{ij} \):

\[ g^{-1}(X'_{ij} \beta) = E(Y_{ij} \mid X_{ij}) = \mu_{ij} = \sum_{l=1}^{n} \phi_{il} g^{-1}(\Delta_{ij} + Z'_{ij} \beta^*); \]

this can be accomplished using Brent’s method (Press, et al., 1992).

4 Application: Clinical trial of contracepting women

Next, we illustrate the application of the proposed method to data from the clinical trial of contracepting women introduced in Section 1. Recall the randomized clinical trial compared two doses of a contraceptive: 4 injections of 100 mg or 150 mg of DMPA were given at 90-day intervals. The outcome of interest is a repeated binary response indicating whether or not a woman experienced amenorrhea during follow-up intervals. In this study there was substantial dropout for reasons that were thought likely to be related to the outcome. More than one third of the women dropped out of the trial: 17% dropped out after the 1st 90 day interval, 13% dropped out after the 2nd 90 day interval, and 7% dropped out after the 3rd 90 day interval.

We considered the following logistic model for the probability of amenorrhea,

\[ \logit E(Y_{ij} \mid X_{ij}) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t \times \text{dose}_i + \beta_4 t^2 \times \text{dose}_i, \]

where \( t = 1, 2, 3, 4 \) represents time elapsed (in terms of 90-day intervals) and \( \text{dose}_i = I(150 \text{mg DMPA}) \). We estimated the vector of treatment-specific dropout probabilities, \( \phi_{\text{dose}_i} \), non-parametrically. Finally, to complete specification of the marginally-specified pattern mixture model we considered three assumptions for dropout:

(a) “next dropout pattern”: For \( j > k \), \( \logit E(Y_{ij} \mid X_{ij}, D_i = k) = \logit E(Y_{ij} \mid X_{ij}, D_i = j) \).

(b) “dropout trend”: \( \logit E(Y_{ij} \mid X_{ij}, D_i = k) - \logit E(Y_{ij} \mid X_{ij}, D_i = j) = \alpha_{(\text{dose})} (k - j) \).

(c) “complete-case contrast”: \( \logit E(Y_{ij} \mid X_{ij}, D_i = k) - \logit E(Y_{ik} \mid X_{ik}, D_i = k) = \logit E(Y_{ij} \mid X_{ij}, D_i = 4) - \logit E(Y_{ik} \mid X_{ik}, D_i = 4) \).
In (a), the mean response at any occasion following dropout is tied to the corresponding mean for those who dropout at the subsequent occasion. In (b), a linear trend in dropout time is assumed for the mean response at any occasion; thus, the mean response at any occasion following dropout is extrapolated from the observable trend across dropout patterns. Note that in both (a) and (b), there is an implicit assumption that those who dropout early in the study are more similar to those who dropout soon after than to those who dropout later or complete the study. In contrast, (c) assumes that any longitudinal trend in the mean response following dropout, relative to the mean prior to dropout, is similar to the corresponding trend for the study completers. Assumptions (a)-(c) were chosen for illustrative purposes only; in a true sensitivity analysis, the right hand side of equations (a)-(c) might include an additional sensitivity parameter whose value is varied across a plausible range. Ideally, any assumptions made should be guided by subject-matter considerations.

Table 1. Estimated marginal probabilities of amenorrhea.

<table>
<thead>
<tr>
<th>Group</th>
<th>Dropout Model</th>
<th>Visit 1</th>
<th>Visit 2</th>
<th>Visit 3</th>
<th>Visit 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>100mg</td>
<td>MCAR</td>
<td>0.174</td>
<td>0.273</td>
<td>0.387</td>
<td>0.500</td>
</tr>
<tr>
<td></td>
<td>Next dropout pattern</td>
<td>0.172</td>
<td>0.288</td>
<td>0.423</td>
<td>0.551</td>
</tr>
<tr>
<td></td>
<td>Dropout trend</td>
<td>0.176</td>
<td>0.295</td>
<td>0.422</td>
<td>0.532</td>
</tr>
<tr>
<td></td>
<td>Complete-case contrast</td>
<td>0.172</td>
<td>0.286</td>
<td>0.415</td>
<td>0.536</td>
</tr>
<tr>
<td>150mg</td>
<td>MCAR</td>
<td>0.211</td>
<td>0.347</td>
<td>0.468</td>
<td>0.547</td>
</tr>
<tr>
<td></td>
<td>Next dropout pattern</td>
<td>0.212</td>
<td>0.369</td>
<td>0.508</td>
<td>0.596</td>
</tr>
<tr>
<td></td>
<td>Dropout trend</td>
<td>0.214</td>
<td>0.376</td>
<td>0.509</td>
<td>0.583</td>
</tr>
<tr>
<td></td>
<td>Complete-case contrast</td>
<td>0.212</td>
<td>0.374</td>
<td>0.521</td>
<td>0.618</td>
</tr>
</tbody>
</table>

The estimated marginal probabilities of amenorrhea at each occasion, under the three dropout assumptions, are presented in Table 1. Also presented in Table 1 are the corresponding estimates under the assumption that dropout is completely at random (MCAR), as a point of reference. In general, the three assumptions concerning non-ignorable dropout shift the marginal probabilities upward (see Figure 1). This suggests that women who dropout have higher risk of amenorrhea. However, treatment differences remain stable and are relatively unaffected by these three assumed dropout processes (see Figure 2).

5 Concluding Remarks

It is an inescapable fact that all methods for handling non-ignorable dropout have to make some unverifiable assumptions; this is true of both selection and pattern mixture models. Selection and pattern mixture models have their own distinct advantages and disadvantages. The proposed marginally-specified pattern mixture models attempt to capitalize on desirable features
Methods for handling dropout

FIGURE 1. Estimated probabilities of amenorrhea under three assumptions of non-ignorable dropout, contrasted to those obtained assuming MCAR: (a) “next dropout pattern”; (b) “dropout trend”; and (c) “complete-case contrast”.

FIGURE 2. Estimated risk difference of amenorrhea under three assumptions of non-ignorable dropout.
of each approach. Specifically, these models circumvent the obvious drawback of pattern mixture models. By construction, the regression parameters in marginally-specified pattern mixture models have “marginal” interpretations. Also, unlike selection models, identifiability restrictions are readily established. Finally, estimation is relatively straightforward using a GEE approach.

Of note, the proposed model is semi-parametric. The avoidance of full distributional assumptions can be seen as advantageous in this setting as it avoids the need to make identifying restrictions on all higher-order moments (Wilkins and Fitzmaurice, 2006). Furthermore, there is often no convenient specification of the joint distribution when the responses are discrete. Finally, we note that the general approach is closely related to “marginally-specified conditional models” developed in the complete data setting (Heagerty and Zeger, 2000).

Acknowledgments: This work is joint with Ken Wilkins and was supported in part by grant GM 29745 from the National Institutes of Health.

References


Methods for handling dropout


Modelling temporal uncertainty in palaeoclimate reconstructions

John Haslett\textsuperscript{1}, Andrew Parnell\textsuperscript{1} and Michael Salter-Townshend\textsuperscript{1}

\textsuperscript{1} Department of Statistics, Trinity College, Dublin 2, Ireland

Abstract: We present a method for reconstructing an aspect of climate over 13,000 years at Sluggan Moss, Northern Ireland. We extend the work of Haslett et al. (2006) to include calibrated radiocarbon ages. The required chronologies are obtained via a change-point regression. We also discuss the $t_\alpha$ long-tailed random walk model and produce predictive climate estimates.

Keywords: Radiocarbon dating; change-point regression; Long-tailed random walk; $t$-process

1 Introduction

Haslett et al. (2006) presented a Bayesian method for the quantitative reconstruction of past climates from data on pollen composition in lake sediment. The essential science is that pollen found in lake sediment reflects ancient vegetation, which in turn reflects the ancient climate. Thus changes in the pollen composition with sediment depth reflect changes in the climate in times past. The key contribution of that paper was an approach that could deal with the many sources of uncertainty involved in this exercise. One such source is chronological uncertainty. Thus, although the data - multivariate counts for pollen of different taxa - correspond to known depths in the sediment, the corresponding dates are at best known imperfectly. This aspect was deferred in Haslett at al. (2006); we provide here an introduction to methods that deal with uncertain chronological information in the context of palaeoclimate reconstruction.

Statistically, the simplest version of the problem may be stated as follows. For each of a number of samples, $n^m$ modern and $n^f$ fossil, vectors of compositional data $p^m = \{p^m_j; j = 1, \ldots, n^m\}$ and $p^f = \{p^f_j; j = 1, \ldots, n^f\}$, are available for study; these are often referred to as ‘pollen assemblages’ or ‘pollen spectra’. For the modern sample, vectors of climate data $c^m = \{c^m_j; j = 1, \ldots, n^m\}$ are also available as covariates; the climate values $\{c^f_i; i = 1, \ldots, n^f\}$ for the fossil sites are missing. In our work, $n^m = 7815$ samples from Europe and N. America are available; the pollen vectors are of length 14 (including an ‘Other’ category). Multivariate climate is two dimensional.
in our work; in the present paper our focus is on one aspect of climate, the length of the growing season, $GDD5$ (growing degree days above 5°C); the other is $MTCO$ (the mean temperature of the coldest month). The objective is to estimate the missing values and thus to reconstruct the prehistoric climate. Haslett et al. (2006) present a Bayesian model for $\pi(cf|pf, cm, pm)$, which we write as $\pi(cf|data)$, the distribution of fossil climate given the data.

In the current study the $pf$ arise from $n' = 115$ samples at Sluggan Moss in Northern Ireland; thus $cf$ is a vector of 115 $GDD5$ values. Figure 3 shows two different reconstructions of $GDD5$ over the past 13,000 years cal BP (calendar years before present) or so. It is the purpose of this paper to discuss the differences between these. Both show that there appears to have been a rapid rise in $GDD5$ at circa 10k cal BP corresponding to circa 9k radiocarbon years BP, shown in the left panel. The right panel differs markedly in the certainty that we can assert for reconstructed climates in the period 10k to 11k cal BP. In particular the reconstructions in that period are bimodal; the HPD regions are disjoint; the signal is weak. The differences between these methods concern the way we model the temporal uncertainty in the dating of samples. It is on this aspect of the modelling that we focus in this paper.

1.1 Modelling Issues

From a statistical point of view there are many subtle issues. Firstly, even if the dates for the 115 sediment samples are taken to be known, and interest to lie exclusively in climate reconstructions at those dates, Bayesian methods require that we have a prior distribution for the unknown climate $cf$ at those dates. As this describes an evolving climate, and interest is focussed on climate change, we are particularly interested in joint priors, that is for the entire vector $cf$. An important novel issue in our approach is that the prior addresses the temporal structure manifest in (some degree of) ‘smoothness’ in time; smoothness is discussed in Section 2.3. But even if known without error, the samples, although regular in depth, are irregular in time. In this paper we are concerned with reconstructions on a regular grid in time (20 year steps over 13,000 years); sedimentation rates are not constant over this time-scale. Further, there is considerable uncertainty in the dating of the samples.

There are many other statistical issues. For example, the underlying data are pollen counts, and there is much evidence of zero-inflation; see Salter-Townshend and Haslett (2006, this volume). Although here climate is represented by one variable, in general it is multivariate; we are exploring low-dimensional additive models as efficient low-dimensional approximations to high-dimensional climate space. Climate is, in fact, a spatio-temporal process; scientific interest lies in reconstructing the entire European palaeoclimate for the Holocene. In this context single sites like Sluggan Moss are
but proof-of-concept studies. Pollen is but one proxy for past climates; our methods can in principle be used for multiple proxy reconstructions. The assessment of model fit for all these situations is computationally challenging; see Haslett et al. (2006), Bhattacharya (2005) and Bhattacharya and Haslett (2006). Our focus in this paper is on temporal smoothness and in particular temporal uncertainty.

In the next section we summarise the statistical methodology, focussing on the temporal smoothness of pre-historical climate; greater detail is given in Haslett et al. (2006). Subsequently we focus on the issue of temporal uncertainty.

2 Climate Reconstruction

One key advance of Haslett et al. (2006) is the recognition, within the statistical model, of the fact that climate is a stochastic process in time. Heretofore, climate has been reconstructed separately for each pollen sample in isolation; indeed this is the how we introduce the problem in section 2.1 below. In fact, we have some knowledge of the temporal aspects of climate; samples that are close in time might reasonably be expected to have similar climates (see section 2.3). We outline the current modelling approach; we then discuss the implications of temporal uncertainty.

The overall model may be thought of in terms of distinct Monte Carlo modules. One module concerns the response surfaces, which we discuss briefly; a second involves temporal smoothness where we propose an improvement on Haslett et al. (2006); the third incorporates temporal uncertainty, on which we give more detail.

2.1 Response surfaces

We summarise here the first stage in this computationally intensive task. At this stage in the discussion it is useful the think of reconstructions as being undertaken point-wise; thus we are concerned with \( \pi(c^f_i|p^m, R) \).

The task here involves modelling, and then using inversely, ‘response surfaces’ \( R \) which describe how multivariate pollen composition \( p \) changes smoothly in two-dimensional climate space. Specifically we define a forward model \( \pi(p|c, R) \); we refer the reader to Haslett et al. (2006) and to Salter-Townshend and Haslett (2006, this volume), for technical details.

The response surfaces \( R \) may be thought of as parameters specifying the forward model. There are approximately 10,000 elements in \( R \), comprising 14 smooth surfaces in two-dimensional climate space.

It is sufficient to say here that the two-stage process presented in Haslett et al. (2006) can be written as

\[
\pi(c^f_i|p^f_i, c^m, p^m) \propto \int \pi(p^f_i|c^f_i, R)\pi(R|c^m, p^m) dR \pi(c^f_i).
\] (1)
The first stage in the implementation is the sampling from \( \pi(R|c^m, p^m) \). The second is the sampling, for a given \( R \), say, from \( \pi(c_f^i|p_f^i, R)\pi(c_f^i) \). Averaging over random replications of \( \{ R_k: k = 1, \ldots, N_R \} \) approximates the integration.

The climate prediction thus obtained via the response surfaces are noisy, producing multiple modes in regions of climate-space. These extra modes occur because the response surfaces are non-monotonic. For example, the pollen used to link climate to a particular taxon arises from many species; different species may be productive in surprisingly diverse climates and thus an abundance of pollen from one taxon may point to several, possibly diverse, climates.

### 2.2 Climate Histories

In Haslett et al. (2006) climate has two dimensions, \( MT CO \) and \( GDD5 \). For the purposes of this paper, it is sufficient to regard the above as an algorithm which delivers vectors, of length 115, of \( GDD5 \) and \( MT CO \) values, all of which are equally likely under the model, given the data. We refer to this as a two-dimensional ‘climate history’. In the next section we outline a method giving preference to ‘smooth’ histories. For simplicity here, we focus solely on \( GDD5 \). Climate histories obtained from the procedure above may exhibit considerable changes in \( GDD5 \) from one time point to the next. However, we have reason to believe that climate changes smoothly in time, in the sense that large climate changes in small time intervals are improbable, though not unknown. We may thus borrow strength from sediment samples at similar depths, for the climate was probably similar. We formalise this below.

The temporal smoothness module concerns sampling climate vectors \( c_f^i = \{ c_f^i; i = 1, \ldots, 115 \} \) from the joint posterior distribution \( \pi(c_f^i|p_f^i, c^m, p^m, d_f^i) \propto \pi(p_f^i|c_f^i, c^m, p^m, d_f^i)\pi(c_f^i) \); the prior \( \pi(c_f^i) \) is of particular importance in this paper. Each element \( c_f^i \) of \( c_f^i \) corresponds to a depth \( d_i \). At this stage in the discussion, these may be considered as corresponding to known times \( \theta_f^i = \theta_f^i(d_i) \) in the past; we present a method for smoothing in section 2.3. Subsequently we consider the relaxation to uncertain knowledge of these times given the depths \( d_i \).

All the histories so generated are jointly consistent (in a probabilistic sense) with the observed data; that is, as entire histories rather than as a series of independent reconstructions at different points in time. Some histories, and some details of those histories, recur more frequently than others; see Figure 3. The role of the prior \( \pi(c_f^i) \) is of great importance here. The independence model \( \pi(c_f^i) = \prod_i \pi(c_f^i) \) implicit in section 2.1 places no joint requirement on the histories; they are a series of independent reconstructions.

The prior we adopt below for \( \pi(c_f^i) \) reflects the temporal smoothness we impose on the reconstructed climate histories. Haslett et al. (2006) discuss some alternatives. To choose an appropriate prior we need to step back from
the details of reconstructions from pollen over a relatively short period of
geological time to that of the past hundred thousand years.

2.3 Climate Smoothness

One such source of information on the variability in climate in this time
scale is that of ice core data from Greenland. We make use of the GISP2
ice core, from which $\delta^{18}O$ has been measured (a proxy for temperature)
over a period longer than 100,000 years. The laminated sections of the core
allow a clear and precise chronology. A plot of the GISP2 ice core, and
a Normal QQ-plot on the associated first differences, is given in Figure
1. It is clear that there have been wide fluctuations in climate, and that
the last 10,000 years has seen an unprecedented period of stability. The
first differences, when examined, show overdispersion compared with the
Normal distribution. Further analysis suggests that the $t_{\nu}$ distribution is a
good description; $\nu = 8$ seems adequate. We use this analysis to motivate
a prior for temporal smoothness. We remark that the results are fairly
insensitive to this particular choice of $\nu$, although the extremes $\nu = 1$
(Cauchy) and $\nu = \infty$ (Normal) are respectively under-constrained and
over-smooth. We subsequently discuss alternatives to this model.

Specifically we follow Haslett et al. (2006), adopting as prior a model
motivated by a random walk in which climate changes are independent re-
alisations of $t_{\nu}$ random variables on a regular temporal grid of 20 years.

Formally, for climate $c_f^i = c_f(\theta_i)$, we write

$$
\pi(c_f) = \prod_i \pi(c_f^i | c_f^{i-1}) = \prod_i t_{\nu_i} \left( \frac{c_f^i - c_f^{i-1}}{\rho_i (\theta_i - \theta_{i-1})^2} \right).
$$

Here $t_{\nu_i}$ indicates the $t$ distribution with $\nu_i$ degrees of freedom, and $\rho_i = h_i \rho$
is related to a constant $\rho$ of the process which quantifies the smoothness.
The reader is directed to the Appendix for details.

3 Temporal uncertainty in climate reconstruction

In the discussion above, we have taken the dates $\theta_i = \theta_f(d_i)$ to be known
functions of depth. This is far from the truth. The process of dating samples
of sediment is prone to considerable uncertainty. It is to this that we now
turn.

Climate reconstructions in Haslett et al. (2006) were presented in radio-
carbon years (often called $^{14}C$ years) which differ substantially from cal-
endar years (cal years). Furthermore, the transformation from $^{14}C$ years
to calendar years, a process known as calibration, introduces complex er-
ror structures. We present a method for taking account of such structures
via Monte Carlo simulation. The use of stochastic dates introduces further
complications in climate reconstructions. In particular, we change our focus, now seeking \( \pi(c_f^G|\text{data}) \) on a regular lattice \( G \) of calendar years. Most importantly, we restrict our grid to be of resolution 20 years, parallel to that of the ice core data. This construction allows the assumed \( t_8 \) distribution to consistently apply between each grid step. This is discussed further in section 4.

Writing \( c_f^\theta \) for the climates at date \( \theta \) for which there are data, we note that \( \pi(c_f^G, c_f^\theta|\text{data}) = \pi(c_f^G|c_f^\theta)\pi(c_f^\theta|\text{data}) \). Thus:

\[
\pi(c_f^G, c_f^\theta|\text{data}) = \int_\theta \pi(c_f^G|c_f^\theta)\pi(c_f^\theta|\text{data})\pi(\theta|d) d\theta
\]  

(3)

where the stochastic process \( \theta \) denotes the true calendar dates corresponding to the depths \( d \), based on the radiocarbon dating. The Monte Carlo algorithm now has a third stage: given a sample of dates \( \theta \) (from depths \( d \)), and the sampling of climates \( c_f^\theta \) at those dates, we now draw a sample of climates \( c_f^G \). A special case is when we constrain the unknown dates: \( \theta \in G \). Below we discuss the uncertainties in \( \theta \). In section 4 we discuss the implications for palaeoclimate reconstruction using stochastic dating.

3.1 Radiocarbon dating

The scientific details of radiocarbon dating are well-described in Bowman (1990) and Aitken (1994). In short, the variable production of \(^{14}\)C (radiocarbon) in the upper atmosphere means that radiocarbon ages do not correspond precisely with calendar ages. However, estimates of this variable production are available in the form of the calibration curve, derived by Reimer et al. (2005). The calibration curve maps radiocarbon years onto
the calendar timescale. It suffers from its own level of uncertainty as it is created from samples which can be dated imprecisely via two methods (one of which must be radiocarbon). The nature of variable $^{14}$C production means that the calibration curve is highly non-monotonic.

The process of calibrating a radiocarbon age is a statistical issue and is dealt with in Buck et al. (1996) and Buck and Millard (2004). In short, the radiocarbon dating laboratory will return the age of a sample in radiocarbon years with an estimate of its associated standard error. The age is usually created from a large sample of Poisson count data; the Normal approximation is assumed appropriate. It is conventional to express the radiocarbon determination in the form $y \pm \sigma$. We define a parameter, $\theta$, to represent the calendar age of the sample. Prior information about this value may be available. We further consider a likelihood model for the radiocarbon determination:

$$ y \mid r(\theta), \sigma \sim N \left( r(\theta), \sigma^2 \right) ,$$  

where $r(\theta)$ is the calibration curve. This is also given a distribution:

$$ r(\theta) \sim N(\mu(\theta), \sigma^2(\theta)) ,$$  

where $\mu(\theta)$ is a piece-wise curve joining known points on the calibration curve, and $\sigma^2(\theta)$ is a quadratic function of the known data points’ variances. A number of ways have been proposed for estimating $\sigma^2(\theta)$; see, for example, Christen (1994).

Now, given the above likelihood and a prior distribution for $\theta$, a posterior distribution of the calendar age can be created. The nature of the calibration curve means that the resulting posterior density can be, on occasion, highly multi-modal.

### 3.2 Radiocarbon dating of sediment cores

The situation when dealing with radiocarbon determinations from lake sediment cores is slightly different. Now, we have a number of different radiocarbon determinations taken at different depths. We require a posterior date distribution for each depth at which fossil pollen has been collected. Thus we have extra data in the form of depths at which each determination is calculated, and prior information for each calendar date as each must be younger than that below it, and older than that above it. We follow the method of Blaauw and Christen (2004), which uses change-point regression. We now have radiocarbon determinations $(y_1, \ldots, y_n)$ with standard errors $(\sigma_1, \ldots, \sigma_n)$ observed at depths $(d_1, \ldots, d_n)$. We define a number of change-points in sedimentation rate at different depths $(u_1, \ldots, u_k)$, which lie at calendar dates $(\phi_1, \ldots, \phi_k)$. These parameters are used in a change-point regression with the relationship:

$$ \theta(d) = \phi_i + \alpha_i(d - u_i-1) , \text{ for } u_{i-1} < d \leq u_i ,$$  

(6)
where $\alpha$ is a measure of the sedimentation rate. The likelihood is now given as:

$$y_j | \ldots \sim N \left( \mu(\theta(d)), \sigma^2(\theta(d)) + \sigma^2_j \right).$$

(7)

Blaauw and Christen (2004) further extend their model by considering the possibility that some radiocarbon determinations may be incorrect. These extra details are not discussed here. The final model will give a probability distribution for depth (ie each layer of fossil pollen). It is realisations from these posterior densities which we use to reconstruct climate.

3.3 The Sluggan Moss chronology

The model of Blaauw and Christen (2004) was applied to the Sluggan Moss core to produce an age-depth chronology. The core has been dated at 32 different depths, ranging from 44.5cm to 518cm. However, fossil pollen has been collected at 115 layers ranging from 2cm to 520cm. Thus a certain degree of extrapolation is required, outside the range of the radiocarbon determinations. An initial choice is required in setting the number of change-points; after a number of trials this was set at 1. The model was run using the Bpeat software as described in Blaauw and Christen (2004) for several million iterations. 95% highest posterior density estimates for each required depth are shown in Figure 2. The sedimentation rate can be seen to increase slightly at around 10k years cal BP. Although, the uncertainty appears small, the density plots show that, in places, this can be of the order of a few hundred years.

4 Reconstruction of $GDD$ for Sluggan Moss

In the case of known dates, each returned fossil climate can be plotted against age. After smoothing, it is trivial to produce point-wise density estimates at each known date. These, when placed adjacently, form neat summaries of former climate, see the left panel of Figure 5. We now extend this method to discuss the use of the constructed radiocarbon chronologies in the climate reconstruction.

4.1 Utilizing temporal uncertainty

The set of age-depth posterior distributions created above are made available to the climate reconstruction program so as to take account of the temporal uncertainty. Fossil climates are updated via MCMC through the $t_s$ random walk.

We are now seeking $\pi(e^f(\theta_G) | data)$ on a regular lattice, $\theta_G$, defined every 20 calendar years. We enforce the continuing $t_s$ structure by only examining
chronological densities at each point on the $\theta_G$ grid. We may represent this as:

$$\pi(c^f(\theta_G)|\text{data}) = \int_\theta \pi(c^f(\theta_G)|\theta_G, \text{data})\pi(\theta_G|\text{d})d\theta$$  \hspace{1cm} (8)

where the stochastic process $\theta_G$ denotes our chronology (from above) corresponding to the depths $d$, based on the radiocarbon dating.

4.2 Simulations on a climate grid

We sample from the posterior chronology distribution to obtain a set of dates, $\theta_G$, and run these through the $t$ random walk to obtain a set of appropriate climates. The random walk aspect of the model is relatively fast, and the posterior date densities are pre-calculated using the chronology software. Thus this aspect of the model is relatively easily implemented. The model can be thought of as being run in two parts. The first part is run only at grid points at which fossil pollen exists. We partition the set $c^f(\theta_G)$ into the part which contains fossil pollen, $c^f(\theta_P)$, and that which does not, $c^f(\theta_P)$. We make use of the kurtosis approximation in section 6 when updating the random walk for periods longer than 20 years. In practice, the approximation forces smoother climate reconstructions than that of Haslett et al (2006) as longer time gaps require $t$-distributions with greater degrees of freedom; consequently the random walks become thinner-tailed.
The second step uses a conditional approximation to fill in those parts of $c^f(\theta_G)$ which do not correspond to fossil pollen data ($c^f(\theta_P)$). Specifically, we require the predictive distribution of $\pi(c^f(\theta_P)|c^f(\theta_P), data)$. We define a joint distribution $\pi(c^f(\theta_P), c^f(\theta_P)|data)$ and treat the random walk as a multivariate $t$-process. Such a joint distribution will allow predictive distributions to be formed using any set of user-inputted dates, $\theta_G$. It is not possible to obtain estimates of this for multivariate $t$ in closed form. However, a predictive density can be formed via Monte Carlo simulation by proposing, and accepting or rejecting climate values which are deemed appropriate by the conditional density. The resulting estimates of $c^f$ are obtained at each of the elements of $\theta_G$. The simulations are entirely consistent with the $t_8$ random walk.

4.3 Application

The process adaptation to the random walk was applied to the Sluggan Moss core and its $GDD5$ reconstruction. This is shown in the right panel of Figure 3. We also include posterior densities of $GDD5$ when the radiocarbon dates are not calibrated. The two panels of Figure 3 show a big shift in timing when moving from $^{14}C$ to cal years BP (note also that this shift in calendar years is contemporary with that of the sedimentation rate change shown in Figure 2). It is also clear that the timing of the major climate shift is made more uncertain with the incorporation of the calendar year uncertainties. It is only this right panel which accurately attempts to model and graph the true uncertainty inherent in both climate and age.

![FIGURE 3. GDD5 reconstructions for Sluggan Moss. Left panel uses $^{14}C$ years, right panel utilises full temporal uncertainty via a gridded $t$-process.](image-url)
5 Summary and extensions

It is clear that temporal uncertainty, whilst affecting the climate reconstructions, presents no major theoretical obstacles to the simulation of fossil climate. However, future versions of the climate reconstruction model require a smooth prior for \( c_f \) which is both consistent (in an additive sense) and allows predictive densities to be formed. It is also highly desirable for the prior to specify a covariance matrix; allowing both temporal and, subsequently, spatial aspects to be included in predictions.

We believe future versions of priors for \( c_f \) will come from a family of closed-under-addition stable distributions. Most importantly, we require these to have long-tails. The extremes of the \( t \)-distribution, the Normal and the Cauchy, do not possess all the desirable properties. However, as a first approximation, the gridded \( t_8 \) random walk provides a reasonable climate reconstruction.

6 Appendix: a model for climate smoothness

Consider first a random walk \( y_i = y_{i-1} + \epsilon_i \) on integer time, where the \( \epsilon_i \) are iid with \( t_\nu \) distribution. Then \( \pi(y_i | y_{i-1}) = t_\nu(y_i - y_{i-1}) \). We seek an approximation for \( \pi(y_i | y_{i-k}) \). As this is the distribution of \( \sum_{i=1}^{k} \epsilon_i \) its distribution is asymptotically normal; it is not \( t_\nu \) (unless \( \nu \) is 1 or infinity). However, we can easily match the first three moments of such a distribution. Furthermore, we know (Abramowitz and Stegun, 1964, P948) that its kurtosis is

\[
kurt(\sum_{i=1}^{k} \epsilon_i) = \frac{1}{k} kurt(\epsilon_i) = \frac{6}{k(\nu-4)}.
\]  

But this is the kurtosis of a \( t \)-distribution with \( \nu_k = 4 + k(\nu - 4) \) degrees of freedom. We thus propose the approximation \( \rho_k \sum_{i=1}^{k} \epsilon_i \sim t_{\nu_k} \), where \( \rho_k \) is a scaling factor to ensure that the variance is also correct; it is easy to show that \( \rho_k^2 = (1/k)^\frac{\nu_k(\nu - 2)}{\nu(\nu - 2)} \). Note that \( k \) defines the elapse of time.

We use these arguments to motivate the approximation used in Equation 2 for \( \pi(c_f^i | c_f^{i-1}) \). If the time interval \( t_i - t_{i-1} \) may be written as an integer multiple of the 20 year resolution of the GISP data then \( k = 20^{-1}(t_i - t_{i-1}) \). More generally, providing that \( t_i - t_{i-1} \geq 20 \) we propose to take \( k \) as the integer part of \( t_i - t_{i-1} \geq 20 \). Equation 2 follows.

References


Quantitative assessment of probabilistic forecasts with applications in epidemiology

Leonhard Held1

1 Department of Statistics, Ludwig-Maximilians-University Munich, Ludwigstr. 33, D-80539 Munich, Germany

Keywords: Age-period-cohort model; model choice; Monte Carlo estimation; predictive distribution; predictive $p$-values; proper scoring rules.

1 Introduction

One of the main purposes of statistical modelling is to make forecasts for future observations. Such out-of-sample predictions will usually provide point predictions and suitable measures of the uncertainty associated with them. In a Bayesian framework, the predictive distribution of observables is the key quantity, whereas frequentist-based prediction intervals are typically based on so-called tolerance regions, for a comparison see for example Geisser (1993) or Bernardo and Smith (1994), Appendix B.4.3.

In epidemiological applications, prediction problems are numerous. For example, forecasting future cancer rates based on cancer registry data is a central problem (e.g. Knorr-Held and Rainer, 2001, Bray, 2002, Baker and Bray, 2005, Clements et al., 2005). In infectious disease epidemiology, several model-based methods for outbreak detection have been proposed recently (Kleinman et al., 2004, Held et al., 2005, Held et al., 2006), the ultimate validation method is a comparison of the one-step-ahead predictive distributions with the actually observed counts (Menn, 2006).

Indeed, to avoid overfitting, model validation is ideally based on the quality of out-of-sample predictions. This is reflected in the popular use of methods such as cross-validation, if no temporal order is available in the data, or assessment of one-step-ahead predictions, if the data come in a particular order. However, it is surprising that the quantities usually used to compare a probabilistic prediction with actually observed data take only account of the point prediction, the uncertainty associated with it is often ignored. For example, for cross-validation Hastie, Tibshirani and Freedman (2001) suggest to use the squared difference between the mean of the predictive distribution and the observed value $y_{\text{obs}}$ to compare the predictive performance of different models, i.e. the score

$$S(Y, y_{\text{obs}}) = -(E(Y) - y_{\text{obs}})^2,$$

(1)
is to be maximized, here \( Y \sim f(y|x) \), \( x \) denotes the data we condition on. Alternatively one could consider the score

\[
S(Y, y_{obs}) = -|E(Y) - y_{obs}|. \tag{2}
\]

While (1) or (2) provide perfectly reasonable estimates of suitably defined prediction errors, they both ignore the uncertainty associated with the point prediction \( E(Y) \). In other words, these and similar methods only look at the sharpness of predictions, while the goal of probabilistic forecasting is to maximize sharpness subject to calibration (Gneiting and Raftery, 2005). Calibration refers to the statistical consistency between the distributional forecasts and the observations, usually assessed by the empirical coverage of prediction intervals at various levels. An alternative method is to calculate the probability integral transform (PIT) value (David, 1984), which should be uniformly distributed for a perfect forecast of continuous variables. For a recent review see Gneiting, Balabdaoui and Raftery (2005).

In this paper I will describe methods for model choice and model criticism based on probabilistic predictions of external data. I do not intend to give a comprehensive overview over the literature on this subject, the reader is referred to Gneiting and Raftery (2005) for further references. The methods described will be illustrated through a case study from chronic disease epidemiology. Other areas of application will be sketched in the discussion.

## 2 Scoring rules

Scoring rules (also called scoring functions) are the key measures for the evaluation of probabilistic forecasts, for a recent review see Gneiting and Raftery (2005). Scoring rules assign a numerical score (or reward) based on the predictive distribution for the unknown quantity and of that quantity’s true value, that has later materialised. Scoring rules are usually positively oriented, i.e. the larger, the better. Scoring rules are called proper, if they do not provide any incentive to the forecaster to digress from her true belief (e.g. Bernardo and Smith, 1994). They are called strictly proper if any such digress results in a penalty, i.e. the forecaster is encouraged to quote her true belief rather than any other predictive distribution.

For illustration, the mean squared prediction error score

\[
S(Y, y_{obs}) = -E(Y - y_{obs})^2 = -\text{Var}(Y) - (E(Y) - y_{obs})^2, \tag{3}
\]

comparing the predictive distribution of \( Y \sim f(y|x) \) with the observed value \( y_{obs} \) is not proper: the forecaster will quote the deterministic forecast \( E(Y) \) rather than her true belief \( f(y|x) \) in order to maximize the expected score (assuming \( y_{obs} \sim f(y|x) \)).

An alternative approach proposed in the literature is to use the squared normalized prediction error (or, in the multivariate case, a squared Maha-
lanobis distance, e.g. O’Hagan, 2003) to assess the quality of the predictions, i.e. the score

\[ S(Y, y_{obs}) = -(E(Y) - y_{obs})^2 / \text{Var}(Y). \] (4)

Also this score is not proper, as the forecaster will quote a forecast with mean \( E(Y) \), but with variance going to infinity in order to maximize the expected score. However, the modification

\[ S(Y, y_{obs}) = -(E(Y) - y_{obs})^2 / \text{Var}(Y) - \log(\text{Var}(Y)) \] (5)

is proper if \( Y \) is normal. Indeed, the expected score is maximized if the forecaster quotes the (normal) distribution with mean and variance identical to \( E(Y) \) and \( \text{Var}(Y) \). This score is a special case (ignoring additive constants) of the so-called logarithmic score

\[ \log S(Y, y_{obs}) = \log f(y_{obs} | x) \] (6)

the log predictive density ordinate at the observed value \( y_{obs} \).

The logarithmic score has been criticized being exceedingly sensitive to outliers and extreme events. Additionally, it is not sensitive to distance, as the predictive density \( f(y | x) \) is evaluated only at \( y_{obs} \), all other values of \( f(y | x) \) do not matter. Indeed, Selten (1998) concludes that

\[ \ldots \text{the logarithmic scoring rule is not really recommendable. On the one hand, it is too sensitive with respect to differences between very small probabilities, on the other hand, it is sometimes not sensitive enough, in the sense that in some situations it does not matter whether the truth is near to the prediction or far from it.} \]

A popular proper score which is less sensitive to outliers but sensitive to distance is the so-called continuous ranked probability score

\[ \text{CRPS}(Y, y_{obs}) = - \int_{-\infty}^{\infty} (P(Y \leq t) - 1(y_{obs} \leq t))^2 dt, \]

which is the integral of the celebrated Brier score for binary predictions at all possible thresholds \( t \). Interestingly, this score can be written as

\[ \text{CRPS}(Y, y_{obs}) = \frac{1}{2} \text{E}|Y - Y'| - \text{E}|Y - y_{obs}|, \] (7)

here \( Y \) and \( Y' \) are independent realisations from \( f(y | x) \) (see Gneiting and Raftery, 2005, for references). The continuous ranked probability score can be generalized to the energy score

\[ \text{ES}(Y, y_{obs}) = \frac{1}{2} \text{E}|Y - Y'|^\alpha - \text{E}|Y - y_{obs}|^\alpha, \] (8)
which is strictly proper for all $\alpha \in (0, 2)$ (Gneiting and Raftery, 2005). In the limiting case $\alpha = 2$, however, $\text{ES}(Y, y_{\text{obs}})$ reduces to (1), as
\[
\frac{1}{2}E(Y - Y')^2 = \text{Var}(Y).
\]
In this case the score is proper, but not strictly proper, the latter is clear as the variability of $Y$ is not taken into account. Indeed, (1) does not change if we change the predictive variance $\text{Var}(Y)$. Note the difference between (1) and the mean squared error score (3).

The CRPS and, more generally, the energy score can be generalized to multivariate observations (see Gneiting and Raftery, 2005):
\[
\text{ES}(Y, y_{\text{obs}}) = \frac{1}{2}E||Y - Y'||^\alpha - E||Y - y_{\text{obs}}||^\alpha,
\]
here $||.||$ denotes the Euclidean norm and, as before, $\alpha \in (0, 2)$ is required to ensure strict propriety.

3 Tools for model criticism

Scoring rules can be used to decide between competing models. However, they do not answer the question if a chosen model is consonant with external data. There are several options to address this issue. One possibility is to calculate the probability integral transform (PIT) value, defined as $p = F(y_{\text{obs}})$ where $F$ is a univariate predictive distribution function and $y_{\text{obs}}$ is the actually observed data. If the forecast is perfect and $F$ is continuous, $p$ will be uniformly distributed (David, 1984). In practice one usually examines the histogram of several PIT values for departure from a uniform distribution. If the data are counts as in our application, a correction is possible to ensure uniform histograms for perfect forecasts (Czado et al., 2006).

A less-known alternative approach, that is also applicable to multivariate distributions, is the calculation of predictive $p$-values (Box, 1980). Let $f(y|x)$ be the (now possibly multivariate) predictive density for external data $y$, conditional on observed data $x$. Interest is in measuring the support of the predictive density for the actually observed data $y_{\text{obs}}$ by considering
\[
P\{f(Y|x) \leq f(y_{\text{obs}}|x)|x\},
\]
where $f(Y|x)$ is a function of the random variable $Y \sim f(y|x)$. This proposal dates back to at least to Box (1980), who considered prior predictive $p$-values, without conditioning on observed data $x$.

Interestingly, if $Y \sim f(y|x)$ follows a continuous distribution, (10) viewed as a function of the random variable $y_{\text{obs}} \sim f(y|x)$ is also uniformly distributed. This allows for multivariate model checks as done usually with PIT histograms for univariate forecasts.
Monte Carlo Estimation

Suppose that samples \( y^{(l)}, l = 1, \ldots, L \) from the predictive distribution \( f(y|x) \) are available. Monte-Carlo estimation of the CRPS score (7), the energy score (8) and its multivariate generalization (9) is straightforward. Also, the PIT values can be estimated based on the empirical distribution function of the predictive distribution.

More challenging is the computation of the logarithmic score and Box’s \( p \)-value. Of course, if the functional form of the predictive distribution \( f(y|x) \) is known, then calculation of (6) is trivial, it does not even require samples from \( f(y|x) \). Slightly more involved, but still straightforward is the estimation of (10) in this case. Indeed, extending the result by Wei and Tanner (1990), (10) can be simulation-consistently estimated by

\[
\frac{1}{L} \sum_{l=1}^{L} \mathbb{1}\{f(y^{(l)}|x) \leq f(y_{obs}|x)\}.
\]

The predictive density \( f(y|x) \) is, however, usually unknown, but in many cases the conditional density \( f(y|x, \eta) \) is known, here \( \eta \) are unknown model parameters. Furthermore, also samples \( \eta^{(m)} \) from the posterior distribution \( f(\eta|x), m = 1, \ldots, M \), will be typically available. We can then adopt an approach first described in Gelfand and Smith (1990) to estimate \( f(y|x) \) by averaging the conditional densities \( f(y|x, \eta^{(m)}) \), \( m = 1, \ldots, M \):

\[
\hat{f}(y|x) = \frac{1}{M} \sum_{m=1}^{M} f(y|x, \eta^{(m)}).
\]

This will enable us to estimate the logarithmic score (6), and also Box’s \( p \)-value (10). For more discussion of the latter approach see Held (2004).

Application to age-period-cohort models

Bayesian age-period-cohort models are used increasingly to project cancer incidence and mortality rates. Data from younger age groups (typically age < 30 years) for which rates are low are often excluded from the analysis. However, a recent empirical comparison (Baker and Bray, 2005) of analyses with and without the data from younger age groups suggested that age-specific projections based on full data sets are more accurate for younger age groups, in particular for age group 30-34. Baker and Bray (2005) notice that “the most striking feature is the increased width of the 90 percent [prediction] intervals based on the reduced compared with the full data set” for data on larynx cancer mortality in Hungary. However, to quantify the accuracy of the predictions, the squared normalized prediction error (4) was used. In the case study described in the following, we will investigate
if the same conclusion can be drawn for mortality data on larynx cancer from Germany, for full details see Menn (2006). We will compute the scores discussed in this paper for model comparison and will use PIT values and Box’s p-values for model criticism.

Let \( n_{ij} \) denote the number of persons at risk in age group \( i \) \((i = 1, \ldots, I)\) and year \( j \) \((j = 1, \ldots, J)\). We assume that the number of deaths \( y_{ij} \) from larynx cancer, in age group \( i \) during year \( j \) has a binomial distribution with parameters \( n_{ij} \) and \( \pi_{ij} \), and that the likelihood for the entire data is the corresponding product of binomial terms. If the age group bands were of the same width as the period bands, for example years, then the classical APC model (e.g. Clayton and Schifflers, 1987) could be adopted which decomposes the log odds \( \eta_{ij} = \log\{\pi_{ij}/(1-\pi_{ij})\} \) additively into an overall level \( \mu \), age effects \( \theta_i \), period effects \( \phi_j \) and cohort effects \( \psi_k \),

\[
\eta_{ij} = \mu + \theta_i + \phi_j + \psi_k, \quad (13)
\]

where the cohort index \( k = 1, \ldots, K \) is defined through \( k = I - i + j \) with maximum value \( K = I - 1 + J \).

Bayesian versions of the APC model typically use a flat prior for \( \mu \) and so-called second-order random walks for the age, period or cohort parameters. Such formulations have close connections to nonparametric smoothing methods, such as penalized likelihood or discrete spline smoothing. For example, for the age effects \( \theta \), the second order random walk (RW2) is given by \( \theta_i|\theta_{i-1}, \theta_{i-2} \sim N(2\theta_{i-1} - \theta_{i-2}, \kappa^{-1}) \), \( i = 3, \ldots, I \) with independent uniform priors both for \( \theta_1 \) and \( \theta_2 \). Here \( \kappa \) is a precision parameter which determines the smoothness of the age effects. A highly dispersed but proper gamma distribution is assigned to the hyperparameter \( \kappa \). The same approach is used for the period and cohort effects and the associated hyperparameters.

One advantage of the Bayesian approach is that the uncertainty about the hyperparameters is incorporated in the marginal estimates of \( \theta, \phi \) and \( \psi \). Furthermore, model (13) can easily be extended, if necessary. For example, to account for overdispersion, random effect type parameters can be added in (13). The model will then be

\[
\eta_{ij} = \mu + \theta_i + \phi_j + \psi_k + z_{ij} \quad (14)
\]

with independent Gaussian variables \( z_{ij} \sim N(0, \delta^{-1}) \). A gamma prior will be assigned to the hyperparameter \( \delta \), similarly as to the other hyperparameters.

A problem with the above definition of cohorts is that it is not applicable when age and period are not on the same grid. A way out of this difficulty is to use the same model as in (13) or (14), but with a different definition of the cohorts. Suppose age is given in five years intervals whereas period is given on an annual basis, as in our dataset. Cohorts can then be defined by \( k = 5 \cdot (I - i) + j \) with maximum value \( K = 5 \cdot (I - 1) + J \).
To sample from the implied posterior distribution by MCMC, parametrization and implementation issues are very important in order to ensure a reliable and efficient algorithm. We refer to Besag et al. (1995) and Knorr-Held and Rainer (2001) for details. Given samples from the posterior distribution, prediction of future death probabilities $\pi_{ij} = 1/(1 + \exp(-\eta_{ij}))$, $j = J+1, \ldots, J+T$, is straightforward to implement. Samples from the predictive distributions of $\phi_j$, $j = J+1, \ldots, J+T$ and $\psi_k$, $k = K+1, \ldots, K+T$ are obtained by repeated application of the RW2 model definition. Similarly, samples from $z_{ij}$, $i = 1, \ldots, I$, $j = J+1, \ldots, J+T$ are generated independently from a $N(0, \delta^{-1})$ distribution. Samples from the predictive distribution of $\pi_{ij}$ and finally given through (14), from which samples from $y_{ij}$ can be easily generated.

We have fitted four different models to the mortality data on larynx cancer among males in Germany, 1952-1997. Model 1 and 2 do include data from age below 30 years, while model 3 and 4 do not. Model 1 and 3 do allow for additional overdispersion, while model 2 and 4 do not.

To assess the predictive quality of the different models, we have predicted the mortality counts in the years 1998-2002. This allows us to compare our predictions with the actual observed data using the methods discussed above. For all four models, we looked at the predictions in the 12 age groups with age above 30 years.

Table 1 gives several averaged scores over all $12 \cdot 5 = 60$ observations. The scores (1), (2) and (3) agree with the CRPS (7) and all prefer model 4. Interestingly, the standardized residual score, also used in Baker and Bray (2005) gives very different results. So do the two versions of the logarithmic score, which may be caused by strong sensitivity to outlying observations. Note that the mean predictive variance, the difference between (3) and (1), is 553.0, 401.9, 573.8, and 383.0. Hence the models allowing for overdispersion do have a larger predictive variance on average, as expected. The comparison based on the proper score (7) suggests that there is a preference for the models without allowance for overdispersion (models 2 and 4). Among those, the model not including the young age groups is preferred. This is in contrast to the results in Baker and Bray (2005). The multivariate CRPS over all five years and all 12 age groups are $-155.6$, $-137.4$, $-158.4$, and $-129.3$ for model 1 to 4 respectively, in accordance with the results based on the mean univariate CRPS. We can also calculate the multivariate CRPS for each age group, which allows for a more detailed assessment of the predictive quality of the different models, see Figure 1. For males, the worst fit can be seen for the age groups between 60 and 64. Interestingly, major differences in the scores between the models can be only seen in the mid age groups (age between 50 and 70). We now turn to model criticism. The histograms of the corrected PIT values for each model are shown in Figure 2. The histograms tend to be U-shaped and there is no clear indication which model is closer to uniformity. However, note that the individual PIT values are dependent, especially within fixed
TABLE 1. Averaged scores for the different models.

<table>
<thead>
<tr>
<th>Score</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>−851.0</td>
<td>−646.2</td>
<td>−874.3</td>
<td>−564.1</td>
</tr>
<tr>
<td>(2)</td>
<td>−20.7</td>
<td>−18.5</td>
<td>−21.1</td>
<td>−17.6</td>
</tr>
<tr>
<td>(3)</td>
<td>−1404</td>
<td>−1048</td>
<td>−1448</td>
<td>−947</td>
</tr>
<tr>
<td>(4)</td>
<td>−1.66</td>
<td>−2.06</td>
<td>−1.67</td>
<td>−2.16</td>
</tr>
<tr>
<td>(5)</td>
<td>−6.75</td>
<td>−6.89</td>
<td>−6.76</td>
<td>−6.93</td>
</tr>
<tr>
<td>(6)</td>
<td>−4.38</td>
<td>−3.97</td>
<td>−6.49</td>
<td>−5.63</td>
</tr>
<tr>
<td>(7)</td>
<td>−14.0</td>
<td>−12.8</td>
<td>−14.3</td>
<td>−12.3</td>
</tr>
</tbody>
</table>

FIGURE 1. Multivariate CRPS of the different models for 12 age groups and 5 years.

age groups, so the number of observations (5·12 = 60) is probably too small to make serious judgements here. We have also computed Box’s p-value for the five-dimensional observations in each 12 age group. Figure 3 shows the corresponding QQ-plots for the four models, which all look like possible realizations if the underlying sample distribution is indeed uniform, given the small sample size. We note the uniformity of Box’s p-value will only hold approximately for discrete data. However, the counts are fairly large, so this approximation can be considered as fairly good.

6 Discussion

Statistical modelling has concentrated in the last decades on fitting models, i.e. estimating parameters of a fixed model (e.g. Davison, 2003). Model choice is rarely done through out-of-sample predictions but usually through information criteria (AIC, BIC) or stepwise selection in a regression context. Validation of statistical methods through external data is the ultimate
model check. The probabilistic nature of statistical forecasts is a key feature that distinguishes statistical methodology from other deterministic-style forecasting techniques, such as deterministic differential equation models, used in infectious disease epidemiology (Anderson and May, 1991) and many other areas of science.

I believe that predictive model criticism, as suggested by Box (1980), will play a central part in the near future. The methods described in this paper are valuable tools to assess the quality of probabilistic forecasts. This trend is already clear visible in areas, where forecasting is of paramount importance, for example in meteorology or in economics, see Selten (1998) who clearly states at the beginning: “Probabilistic theories [...] need to be compared with respect to their predictive success”.

FIGURE 2. Histograms of the PIT-value for the different models
As an example, I have concentrated on an application from chronic disease epidemiology. Another area that would certainly benefit from a more predictive view is longitudinal data analysis (e.g. Diggle et al., 2002). Indeed, whereas forecasting plays a central role in time series analysis, it does not in the analysis of longitudinal data. For example, Box’s multivariate $p$-value could be applied to multivariate forecasts of individual response profiles.

Acknowledgments: Special Thanks to Tilmann Gneiting for helpful discussions and Petra Menn for providing computational assistance.

FIGURE 3. QQ-plot of Box’s $p$-value for the 12 age groups and the different models
References


New methods for including covariates in models for the survival of wild animals

Byron J.T. Morgan

1 NCSE, Institute of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent, CT2 7NF, England.

Abstract: Wild animals are monitors of the environment, as their survival rates reflect effects such as climate change, or new farming practices. In this paper we review the statistical methods used for estimating the survival of wild animals by analysing mark-recapture-recovery data, and we describe the incorporation of various kinds of covariates in models for survival. Several new approaches are described, some of which are classical, and some Bayesian.

Keywords: Conditional analysis; life history strategies; missing values; splines; weather.

1 Introduction and motivation

Wild animal populations are being influenced by human activities to an unprecedented extent because of environmental change and destruction. Some scientists interpret the impact of recent anthropogenic change as the 6th mass extinction because so many species are now regarded as endangered. Thus studies of demographic parameters in animal populations have developed considerably over the last 40 years. A major source of information is data obtained from marked individuals. The use of individually marked animals for scientific study can be traced back to Hans Christian Mortensen, who marked 164 starlings (Sturnus vulgaris) in Denmark in 1899. Web-sites such as

Specialist computer software has ensured rapid technology transfer, so that computer packages such as MARK (www.cnr.colostate.edu/ gwhite/) and M-SURGE (www.cefe.cnrs-mop.fr/wwwbiom/Soft-CR/m-surge.htm) are now widely used for classical inference. Modern Bayesian methods are also currently being developed; see for example Brooks et al (2000). The three main aims in studying marked animals, are to investigate movement and survival, and to estimate abundance. Here we focus on estimating probabilities of annual survival. Data from marked animals are of three main types. In mark-recapture studies the initial marking is followed by

<table>
<thead>
<tr>
<th>Year of release</th>
<th>Number released</th>
<th>Year of first recapture (1981+)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>1981</td>
<td>22</td>
<td>11 2 0 0 0</td>
</tr>
<tr>
<td>1982</td>
<td>60</td>
<td>24 1 0 0 0</td>
</tr>
<tr>
<td>1983</td>
<td>78</td>
<td>34 2 0 0</td>
</tr>
<tr>
<td>1984</td>
<td>80</td>
<td>45 1 2</td>
</tr>
<tr>
<td>1985</td>
<td>88</td>
<td>51 0</td>
</tr>
<tr>
<td>1986</td>
<td>98</td>
<td>52</td>
</tr>
</tbody>
</table>

one or more occasions when live animals may be recaptured or resighted. In mark-recovery studies only the marks of dead animals are recovered. Mark-recapture-recovery (MRR) data result in life histories with, for each animal, a sequence of observations denoting whether or not the animal was observed alive on each occasion, with the record possibly terminating with the discovery of the dead animal.

2 Data and models

2.1 Data

Here we present three small illustrative examples from ornithology.

1. Capture-recapture data. The data of Table 1 are taken from Lebreton et al (1992) and describe a recapture study of dippers (*Cinclus cinclus*). These data result from a small-scale study of adult birds of unknown age. The data may be described by models which involve parameters which are probabilities of annual survival, $\phi$, and recapture, $p$. By making different assumptions regarding these parameters, we produce different models for the data.

2. Ring-recovery data. The example of Table 2, taken from Catchpole et al (1999), results from a national study of lapwings (*Vanellus vanellus*). The data describe the numbers of dead birds found and reported in each year after ringing.

3. MRR data. An illustration is provided in Table 3, corresponding to observations on shags (*Phalacrocorax aristotelis*), observed by Mike Harris on the Isle of May in Scotland. In this case the data take the form of individual life-histories. A 1 indicates that the corresponding bird was captured or recaptured, a 2 indicates when a bird was recovered dead (after which the record for that bird only contains 0s) and a 0 denotes neither a recovery nor a recapture.
TABLE 2. Recovery data for British lapwings ringed shortly after birth during the years 1963 - 1974. Ringing took place before 1 May in each year.

<table>
<thead>
<tr>
<th>Year of ringing</th>
<th>Number ringed</th>
<th>Year of recovery (1900+)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>64</td>
</tr>
<tr>
<td>1963</td>
<td>1147</td>
<td>14</td>
</tr>
<tr>
<td>1964</td>
<td>1285</td>
<td>20</td>
</tr>
<tr>
<td>1965</td>
<td>1106</td>
<td>10</td>
</tr>
<tr>
<td>1966</td>
<td>1615</td>
<td>9</td>
</tr>
<tr>
<td>1967</td>
<td>1618</td>
<td>12</td>
</tr>
<tr>
<td>1968</td>
<td>2120</td>
<td>9</td>
</tr>
<tr>
<td>1969</td>
<td>2003</td>
<td>10</td>
</tr>
<tr>
<td>1970</td>
<td>1963</td>
<td>8</td>
</tr>
<tr>
<td>1971</td>
<td>2463</td>
<td>4</td>
</tr>
<tr>
<td>1972</td>
<td>3092</td>
<td>7</td>
</tr>
<tr>
<td>1973</td>
<td>3442</td>
<td>15</td>
</tr>
</tbody>
</table>

TABLE 3. Illustrations of life-histories for 5 different shags. In the first cohort, both birds are marked at time \( t_1 \), while in the second cohort, the three birds are all marked at time \( t_2 \).

<table>
<thead>
<tr>
<th>Capture/recapture occasions</th>
<th>Cohort</th>
<th>( t_1 )</th>
<th>( t_2 )</th>
<th>( t_3 )</th>
<th>( t_4 )</th>
<th>( t_5 )</th>
<th>( t_6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

2.2 Models for ring-recovery data

We want to estimate annual survival probabilities, \( \phi \), from ring-recovery data, and this involves writing the multinomial probabilities \( \theta_j \) in each row of a recovery table in terms of survival probabilities, which means that we also have to introduce probabilities of dead animals being found and reported dead. We use \( \lambda \) for recovery probabilities. There are many different ways of making this reparameterisation, each resulting in a different model. A striking feature of Table 2 is the relatively large numbers of birds that are found and reported dead in their first year of life. It is a common feature that for wild animals there is appreciable mortality in the first year of life, and so one would include in a model for the data of Table 2, and for similar data, some form of age-dependence in the survival probabilities.
The simplest way of doing this is to have a survival probability $\phi_1$ for animals in their first year of life, and a survival probability $\phi_a$ for all older animals. For such a model, if we have a constant recovery probability the multinomial probabilities for any cohort are:

$$\tilde{\phi}_1 \lambda \quad \phi_1 \tilde{\phi}_a \lambda \quad \phi_1 \phi_a \tilde{\phi}_a \lambda \ldots$$

Note that for any parameter $\eta$ we write $\tilde{\eta}$ for $1 - \eta$. For each cohort we form a separate multinomial likelihood, and then the likelihood for the entire recovery table is simply the product of the multinomial likelihoods for each of the cohorts, which we may write as $L(\phi_1, \phi_a, \lambda)$. This is then maximised to produce the maximum-likelihood estimates of the three parameters.

### 2.3 Models for capture-recapture data

The form of Table 1 is essentially the same as that of Table 2, and the modelling again involves a likelihood which is the product of multinomial likelihoods, one from each cohort. In the case of dippers, we are dealing with birds which are all adults, and so we may suppose that they share the same survival probability, $\phi$. With $p$ denoting the probability of recapture, then for a model with just these two parameters, the multinomial probabilities corresponding to any cohort of marked birds have the form:

$$\phi p \quad \phi^2 \tilde{p}p \quad \phi^3 p^2 \ldots$$

The likelihood, $L(\phi, p)$ can then be formed and maximised.

### 2.4 Models for life-history data

A more interesting model is needed when recovery and recapture data are recorded on the same individuals. Suppression of 1s (2s) in the case-history data results in recovery (recapture) data alone. However, we want to use the combined data, in order to estimate model parameters with the greatest precision. Suppose there are $T$ sampling times, denoted by $t_1, \ldots, t_T$, with recovery information possibly extending further, to $t_K$. Let $C$ denote the number of cohorts of marked animals, where cohort $c$ is defined as all animals first marked at $t_c$. We assume that all animals behave independently, and that, for an animal from cohort $c$, $c = 1, \ldots, C$ and $j = c, \ldots, K - 1$,

$$\phi_{c,j} = \Pr(\text{an animal, alive at } t_j, \text{ survives until } t_{j+1}),$$

$$\lambda_{c,j} = \Pr(\text{an animal which dies in } (t_j, t_{j+1}) \text{ has its death reported})$$

while for $j = c + 1, \ldots, K$,

$$p_{c,j} = \Pr(\text{an animal alive at } t_j \text{ is captured at } t_j).$$
It is shown by Catchpole et al (1998) that we can specify the likelihood in the form,

\[ L \propto C \prod_{c=1}^{C} \left( K - 1 \prod_{j=c}^{T} \{ \alpha_{c,j} \phi_{c,j} \lambda_{c,j} \}^{d_{c,j}} \prod_{j=c}^{T-1} \{ \alpha_{c,j} \chi_{c,j} \}^{\nu_{c,j}} \prod_{j=c}^{T-1} p_{c,j+1} w_{c,j+1} \phi_{c,j+1} \right) . \] (1)

Here the term \( \chi_{c,j} \) is the probability that an animal from cohort \( c \) aged \( j \) is not observed after time \( t_j \), either alive or dead, and it is given by the recursion,

\[ \chi_{c,j} = \phi_{c,j} \lambda_{c,j} \chi_{c,j} + \phi_{c,j} \{ 1 - p_{c,j+1} \chi_{c,j+1} \} , \]

with, \( \chi_{c,K} = 1 \), for \( 1 \leq c \leq C \). Additionally,

\[ \alpha_{c,j} = \begin{cases} 1, & j = c \\ \prod_{s=c}^{j-1} \phi_{c,s} , & c + 1 \leq j \leq K . \end{cases} \]

The likelihood is therefore determined by four sufficient upper triangular matrices, defined below, all relating to animals from cohort \( c \), for \( 1 \leq c \leq C \).

- \( w_{c,j} \) - the number of animals that are recaptured at time \( t_{j+1} \), \( c \leq j \leq T - 1 \);
- \( z_{c,j} \) - the number of animals that are not recaptured at time \( t_{j+1} \) but are encountered later in the study, either alive or dead, \( c \leq j \leq T - 1 \);
- \( d_{c,j} \) - the number of animals that are observed dead in the interval \( [t_j, t_{j+1}) \), \( c \leq j \leq K - 1 \).
- \( v_{c,j} \) - the number of animals that are observed for the last time at time \( t_j \), \( c \leq j \leq T \).

The general formulation given here allows a wide range of complex models to be specified, making different assumptions about how various model parameters may vary with time and age.

2.5 Parameter redundancy

Especially with detailed data, it is tempting to include large numbers of parameters in the models, but sometimes it is not possible to uniquely estimate certain parameters, irrespective of the extent of the data. A model based on a parameter vector \( \theta \) of dimension \( p \) is parameter redundant, with deficiency \( d \), if it can be re-parameterised in terms of a parameter vector of dimension \( p - d \). It is important to be able to detect such parameter redundancy, and symbolic algebra tools are now available to test for this (see Catchpole and Morgan, 1997).
3 Incorporating covariates using logistic regression

Biologists often collect covariates that may influence survival. For example, population density, food availability, genotype and condition of the animal are all factors that may influence survival across species. The North Atlantic Oscillation (NAO) is a measure of the pressure gradient between Stykkisholmur, Iceland and Lisbon; see www.met.rdg.ac.uk/cag/NAO. The average NAO over the winter months has been found to be a good measure of overall winter severity in Europe. The distinction between individual and environmental covariates is an important one.

The use of covariates in MRR studies dates from North and Morgan (1979). In that paper the authors modelled the survival of grey herons (Ardea cinerea), with the annual survival probability, \( \phi_1 \), of birds in their first year of life being a logistic function of the number \( W \) of winter days when the temperature in central England was below freezing: that is,

\[
\text{logit}(\phi_1) \equiv \log \left\{ \frac{\phi_1}{1-\phi_1} \right\} = \beta_0 + \beta_1 W
\]  

where \( \beta_0 \) and \( \beta_1 \) are parameters. Such regressions are valuable, as they both suggest causes for mortality and reduce the number of parameters in models. There are now many examples of the use of covariates in MR work. For instance, Besbeas et al (2002) demonstrate how combining information from different studies may increase the significance of covariates in logistic regressions. Barry et al (2003) have included random effects by the simple addition of a random component to the right-hand-side of Equation (2), and then used a Bayesian analysis. Catchpole et al (2001) analysed MRR data on abalone (Haliotis rubra) and modelled survival as a function of size, estimated from a fitted growth curve.

4 Use of P-splines

Greater flexibility when modelling covariates can be achieved by using splines, which is the approach adopted by Gimenez et al (2006). They consider a nonparametric regression model for the probability that an animal survives from time \( t_i \) to time \( t_{i+1} \) of the form

\[
\text{logit}(\phi_i) = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, T - 1
\]

where \( x_i \) is the value of the covariate for the \( i \)th sampling occasion, \( \varepsilon_i \) are i.i.d \( N(0, \sigma_\varepsilon) \), \( \varepsilon_i \) is independent of \( x_i \) and \( f \) is a smooth function. Here, the random effects \( \{\varepsilon_i\} \) allow us to model the residual variation not described by the covariates alone (Barry et al, 2003). Penalized splines using the truncated polynomial basis were used to model the smooth function:

\[
f(x|\eta) = \beta_0 + \beta_1 x + \ldots + \beta_P x^P + \sum_{k=1}^{K} b_k (x - \kappa_k)_+^P
\]  

where
where \( P \geq 1 \) is an integer, \( \eta = (\beta_1, \ldots, \beta_P, b_1, \ldots, b_K)^T \) is the vector of regression coefficients, \((u)^+ = u^T(\eta \geq 0)\) and \( \kappa_1 < \kappa_2 < \ldots < \kappa_K \) are fixed knots. They used a penalty approach for a fixed number of knots: \( K = \min\{14, 35\} \) and let \( \kappa_k \) be the sample quantiles of the \( x_i \)'s corresponding to probabilities \( k/(K + 1) \). A quadratic penalty is placed on \( b \) which is here the set of jumps in the \( P \)th derivative of \( f(\bullet|\eta) \) so that with Equation (3) we associate the constraint

\[
\begin{align*}
\mathbf{b}^T \mathbf{b} & \leq \lambda 
\end{align*}
\]

where \( \lambda \) is a smoothing parameter. Equations (3) and (4) result in the P-splines approach. P-spline models can be fruitfully expressed as GLMMs, which facilitates their implementation in standard software, and provides a unified framework for generalizations of the nonparametric model.

4.1 An illustration: the effect of sex on survival

The approach is illustrated with data from a 40-year study on individually marked Snow petrels, nesting at Petrels Island, Terre Adélie, from 1963-2002; here \( T = 39 \), and there were 630 males and 640 females. The Southern Oscillation Index (SOI, the counterpart of the NAO for the southern hemisphere) was used as a summary of the overall climate condition (see [www.cru.uea.ac.uk/cru/data/soi.htm](http://www.cru.uea.ac.uk/cru/data/soi.htm)). Gimenez et al (2006) used linear splines \((P = 1)\) but quadratic or even cubic splines could have been used instead, and \( K = 10 \) knots were chosen. The resulting model was as shown below

\[
\begin{align*}
\logit(\phi_{li}) = \beta_0 + \gamma \text{SEX} + \beta_1 \text{SOI}_i + \sum_{k=1}^{10} b_k (\text{SOI}_i - \kappa_k)_+ + \varepsilon_i 
\end{align*}
\]

where \( \phi_{li} \) is the survival probability over the interval \([t_i, t_{i+1}]\) for \( l = \text{male} \) \((\text{SEX} = 0)\) or \( l = \text{female} \) \((\text{SEX} = 1)\) and \( \text{SOI}_i \) denotes the SOI in year \( i \), \( i = 1, \ldots, T \). The random effects \( \{b_k\} \) are independent as well as the \( \{\varepsilon_i\} \). They used the software WinBUGS ([www.mrc-bsu.cam.ac.uk/bugs/](http://www.mrc-bsu.cam.ac.uk/bugs/)). Results not reproduced here suggest that male petrels survive better than females. It also appears that survival is nonlinearly related to the SOI covariate. From a biological point of view, lower values of the SOI may favour access to prey, whereas higher values may improve prey abundance, resulting in the non-linearity found.

5 Sheep and deer: age, population size and individual covariates

Two prime examples in which the age of the animal is an important covariate result from long-term studies of Soay sheep \((Ovis aries)\) and red deer
B.J.T. Morgan  57

(Cervus elaphus). Statistical analyses of data from these studies are given in Catchpole et al (2000), for the sheep, and Catchpole et al (2004) for the deer. In both these applications the regressions are logistic, as in Section 3, and in the papers cited a classical approach was adopted for selection of covariates. As in Section 4.1, sex is an important determinant of survival, but in this case males survive less well than females.

5.1 An age-structured model for the survival of female deer

An example of one of these models is given below, for female deer,

$$
\phi_1(P + N + B), \phi_2, \phi_{3:8}(R), \phi_{9+}(\text{age} + N + R) \mid \nu(P + Y) \mid \lambda(t).
$$

The interpretation is that there is an annual survival probability, $\phi_1$, for animals in their first year of life, and this is logistically regressed on $P$, the population size, and $N$, the average winter NAO, as well as on birth weight $B$. A separate, constant probability of annual survival, $\phi_2$, applies to animals in their second year of life. Animals aged 2–7 share the same annual survival probability $\phi_{3:8}$, which is logistically regressed on the binary covariate $R$, denoting the breeding status of the hind. Animals aged at least 8 years have an annual survival probability $\phi_{9+}$, which is a logistic function of the age of the animal, NAO, and reproductive status. In addition the model allows for an annual dispersal probability $\nu$, which is a logistic function of the population size and a measure $Y$ of the location of the animal. In this model the reporting probability of dead animals is a general time-varying parameter, $\lambda(t)$, which is not related to any of the covariates, and the recapture probability $p = 1$. As an illustration of how this model may be used, Catchpole et al (2004) showed that the probability of death of adult female deer doubles if the deer have given birth in any year.

5.2 Bayesian variable selection for sheep: use of RJMCMC

Variable selection with a range of covariates for each of several model parameters is difficult. One approach is to use reversible jump Markov chain Monte Carlo (RJMCMC) methods, and King et al (2006) have shown how RJMCMC can determine both age-classes and covariate dependence simultaneously. Illustrative results are shown in Table 4 for three constant individual covariates.

6 Time-varying individual covariates: missing data

When individual covariates are present, it is no longer possible to use the expression for the likelihood given in Equation (1), and it is necessary to construct a likelihood by individual (indexed by $i$ in what follows). We now
TABLE 4. The posterior model-averaged estimates of the regression coefficients for the individual covariates for the female sheep, conditional on the survival rate being dependent on the covariate. SD denotes standard deviation. Values that are large relative to the associated SDs are shown in bold. The values for coat type relate to the difference of light to dark coats; while for horn types the differences relate to a third, scurred, horn type.

<table>
<thead>
<tr>
<th>Year of life</th>
<th>Coat type</th>
<th>Horn type</th>
<th>Birth weight</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Light</td>
<td>Polled</td>
<td>Classical</td>
</tr>
<tr>
<td>1</td>
<td>0.332 (0.247)</td>
<td>-0.875 (0.302)</td>
<td>-0.834 (0.315)</td>
</tr>
<tr>
<td>2</td>
<td>0.137 (0.331)</td>
<td>-0.492 (0.350)</td>
<td>-0.449 (0.322)</td>
</tr>
<tr>
<td>3</td>
<td>0.137 (0.331)</td>
<td>-0.492 (0.350)</td>
<td>-0.449 (0.322)</td>
</tr>
<tr>
<td>4</td>
<td>0.193 (0.371)</td>
<td>-0.496 (0.398)</td>
<td>-0.326 (0.414)</td>
</tr>
<tr>
<td>5</td>
<td>-0.055 (0.522)</td>
<td>-0.799 (0.504)</td>
<td>0.883 (1.120)</td>
</tr>
<tr>
<td>6</td>
<td>0.231 (0.948)</td>
<td>-0.838 (0.520)</td>
<td>0.694 (0.993)</td>
</tr>
<tr>
<td>7</td>
<td>0.231 (0.948)</td>
<td>-0.838 (0.520)</td>
<td>0.694 (0.993)</td>
</tr>
<tr>
<td>8</td>
<td>-1.052 (0.410)</td>
<td>0.087 (0.434)</td>
<td>0.156 (0.423)</td>
</tr>
<tr>
<td>9</td>
<td>-1.052 (0.410)</td>
<td>0.087 (0.434)</td>
<td>0.156 (0.423)</td>
</tr>
<tr>
<td>≥10</td>
<td>-1.052 (0.410)</td>
<td>0.087 (0.434)</td>
<td>0.156 (0.423)</td>
</tr>
</tbody>
</table>

present a new way of doing this, presented by Catchpole et al (2006). We denote by $h_{i,r}$ the life-history data entry at time $t_r$, using the notation

$$h_{i,r} = \begin{cases} 
0, & \text{if the } i\text{th animal is not seen at } t_r, \\
1, & \text{if the } i\text{th animal is seen alive at } t_r, \\
2, & \text{if the } i\text{th animal is found dead in } (t_{r-1}, t_r). 
\end{cases}$$

We then define the transition probabilities

$$\pi_{i,r}(a, b) = \Pr(h_{i,r+1} = b \mid h_{i,r} = a).$$

In addition, we define

$$\chi_{i,r,s} = \Pr(\text{not found, alive or dead, from } t_{r+1} \text{ to } t_s \text{ inclusive } \mid \text{ alive at } t_r),$$

for $s = r+1, \ldots, k$, with $\chi_{i,r,r} = 1$. Then we have the recurrence relation

$$\chi_{i,r,s} = (1-\phi_{i,r})(1-\lambda_{i,r})+\phi_{i,s}(1-p_{i,s})\chi_{i,r+1,s}, \quad c_i \leq r < s \leq k-1, \quad (6)$$

which enables $\chi_{i,r,s}$ to be calculated. Here $c_i$ is the time of first capture of the $i$th animal. It is simple to derive the following expressions for $\pi_{i,r}(a,b)$, in an obvious, individual-based notation. In each of the cases where $a = 0$, $\ell$ denotes the occasion on which the animal was last seen alive. Then

$$\pi_{i,r}(0,0) = \chi_{i,\ell,r+1}/\chi_{i,\ell,r}$$

$$\pi_{i,r}(0,1) = \prod_{s=\ell}^{r-1} \phi_{i,s}(1-p_{i,s}) \times \phi_{i,r}p_{i,r}/\chi_{i,\ell,r}$$

$$\pi_{i,r}(0,2) = \prod_{s=\ell}^{r-1} \phi_{i,s}(1-p_{i,s}) \times (1-\phi_{i,r})\lambda_{i,r}/\chi_{i,\ell,r}$$
\[ \pi_{i,r}(1,0) = \chi_{i,r,r+1} \]
\[ \pi_{i,r}(1,1) = \phi_{i,r}p_{i,r} \]
\[ \pi_{i,r}(1,2) = (1 - \phi_{i,r})\lambda_{i,r}, \]

and \( \pi_{i,r}(2,0) = 1 \), with the first of the above equations replaced by \( \pi_{i,r}(0,0) = 1 \) after an animal has been recovered dead. These probabilities enable the likelihood to be constructed directly from the life-history data on each of the \( N \) animals, in the form

\[ L = \pi_{i,r}(1,0)\pi_{i,r}(0,1)\pi_{i,r}(1,1)\pi_{i,r}(1,0)\pi_{i,r}(0,0). \]

where \( w_{i,r}(a,b) \) is an indicator statistic, equal to 1 if \( h_{i,r} = a \) and \( h_{i,r+1} = b \) and equal to 0 otherwise. We can use this expression of the likelihood to deal simply with missing individual covariates as illustrated below. Consider an example with \( T = 6 \) MRR occasions, where a particular animal has the history \((1,0,1,0,0,0)\), and the ordering from left to right corresponds to increasing time. Dropping the subscript \( i \) for clarity, the traditional likelihood contribution for this animal is

\[ \phi_1(1-p_1)\phi_2p_2\phi_3p_3\chi_4, \]

whereas the contribution to the likelihood (7) is

\[ \pi_1(1,0)\pi_2(0,1)\pi_3(1,1)\pi_4(1,0)\pi_5(0,0). \]

Now if each \( \phi \) depends on a current individual covariate, then \( \phi_2 \) and \( \phi_5 \) (and hence \( \chi_4 \)) are unknown, since the animal is not seen on occasions 2 and 5, and so the likelihood (8) is unknown. In practice this may be dealt with by deleting this animal from the likelihood. Similarly, in (9), \( \pi_2(0,1) \) and \( \pi_5(0,0) \) are unknown. But because (9) is based on conditional (transition) probabilities, using (9) we simply delete these transitions from the likelihood, to leave

\[ \pi_1(1,0)\pi_3(1,1)\pi_4(1,0). \]

Thus this approach uses more of the available information than the standard likelihood when there are missing individual covariates.

### 7 Discussion and future work

- The approach of Section 6 is compared with alternative ways of dealing with time-varying individual covariates in Catchpole et al (2006). For example, other conditional analysis methods are attractive to ecologists because they may allow them to use GLMs to incorporate covariates; relevant discussion is provided by Catchpole et al (2004). The conditional method of Section 6 is recommended. Current work by S. Bonner compares this approach with the alternative of imputing missing covariates based on a diffusion model.
• The RJMCMC approach to dealing with covariates in section 5.2 is complex to carry out, and full details are given in King et al (2006). However it allowed a more extensive exploration of covariate selection than was possible in classical analysis. For instance, the finding, in Table 4, that coat colour affects the survival of old sheep was new. Recent work by K. Moyes has established that life-time breeding history is also important for female deer survival, and current research investigates life history strategies for these animals.

• New work by Gimenez and co-workers has successfully applied the spline approach of section 4 to the case of individual covariates.

• A difficulty with environmental covariates, discussed in section 3, can arise if one is dealing with national recovery data, as in Table 2. Current research with D. Brown and I.T. Jolliffe involves interpolating weather maps to produce estimates of local weather which may be used in MRR models.

Acknowledgments: This paper describes joint research with a wide range of statisticians and ecologists. I gratefully acknowledge the collaboration of: Takis Besbeas, Simon Bonner, Daniel Brown, Ted Catchpole, Tim Coulson, Olivier Gimenez, Mike Harris, Ian Jolliffe, Ruth King, Kelly Moyes and Giacomo Tavecchia. The research described in this paper has been supported by grants from the Leverhulme Trust, Marie Curie Framework VI, EPSRC, BBSRC and NERC

References


Part II

Contributed Papers
A Full Bayesian Approach to Physical Maps of Circular Genomes - Application to the Genome of O. oeni strain GM

Maria Antónia Amaral Turkman¹, Luzia Gonçalves² and Libia Zé-Zé¹

¹ Faculdade de Ciencias, Campo Grande, Universidade de Lisboa, 1749-016, Lisboa, Portugal
² Instituto de Higiene e Medicina Tropical, Lisboa, Portugal

Abstract: The first step in bacterial physical mapping typically involves the separate use of at least two rare cutting restriction enzymes to cut the chromosome in a manageable number of fragments (ideally, up to 30), called macrorestriction fragments, that are after separated by PFGE and their sizes estimated by linear interpolation with two flanking size standards. To help the assignment of the restriction fragments, with each other and in the chromosome, several different approaches can be pursued. The advantages of a combined statistical approach during map construction in determining the overlapping probabilities of macrorestriction fragments and directing experimental procedures is already recognized (Zé-Zé et al., 2006).

The overlap detection between fragment pairs is an important step to help in physical mapping. The statistical analysis does not provide the identification of the "true" overlaps, but it can help in simplifying the combinatorial puzzle where pieces are restriction fragments.

In this work we propose a full Bayesian approach towards the characterization of some steps while constructing physical maps of chromosomes. In Gonçalves et al. (2005), the problem is approached as a Bayesian hypothesis test, taking into account the type of prior information based on the length of the fragments and on the size and shape of the chromosome (linear or circular). There, it was assumed that the length of the DNA fragments were constants. However, the measurement of the fragment length can be subject to laboratorial errors. Hence, in this work, we take into account the random character of these measurements to fully analyse the problem from a Bayesian view point.

Keywords: Physical mapping; Overlap probabilities; Bayesian methodology.

1 Prior information and hybridization data

The lengths of fragments can be very informative to identify the overlapping relationship between two fragments. Consider a circular genome and let $N$ be the genome length in kilobases (kb). Consider two fragments $A_i$ and $B_j$, 

---

65
obtained applying enzymes A and B, separately, with fixed lengths $L_i$ and $M_j$. For example, if we have two fragments of a circular genome $A_i$ and $B_j$, such that $L_i + M_j > N$, then they are obligatory overlapped. However, in general, $L_i + M_j \leq N$. The overlapping relationship between two fragments obtained by applying two different restriction enzymes, separately, is classified as nonoverlapping, partial overlapping and total overlapping. In mathematical terms, we can describe it using a variable $\theta_{ij}$ which represents the fraction of overlap ($0 \leq \theta_{ij} \leq 1$). If $A_i$ and $B_j$ are nonoverlapped, then $\theta_{ij} = 0$. If $A_i$ and $B_j$ are totally overlapped, that is, the smaller is contained in the larger, then $\theta_{ij} = 1$ and if fragments are partially overlapped then $0 < \theta_{ij} < 1$.

In a Bayesian perspective, Gonçalves et al. (2005) proposed a mixed prior probability distribution for $\theta_{ij}$. The values 0 and 1 of $\theta_{ij}$ are pointed out and they represent non-overlap and total overlap, respectively. Let $p_{ij}^0$, $p_{ij}^1$ and $p_{ij}^2 = 1 - p_{ij}^0 - p_{ij}^1$ be the probabilities of non-overlap and partial overlap and total overlap, respectively. For $\theta_{ij} \in [0,1]$ we propose a uniform distribution.

Gonçalves et al. (2005) derived expressions for prior probabilities of those events, taking into account DNA fragment lengths and under the assumption that the left-hand endpoints of the two restriction fragments are independent random variables, each of which with a uniform distribution along a circular genome. Prior information, based on fragment lengths, can be combined with hybridization data via Bayes’ theorem, in order to evaluate corresponding posterior probabilities (represented by $p_{ij}^{*0}$, $p_{ij}^{*1}$ and $p_{ij}^{*2}$).

Following previous works developed in the context of clone mapping (e.g. Nelson and Speed, 1994), we consider a set of $m$ probes which defines a comparison vector $D = [d_1, d_2, ..., d_m]$ where each $d_s$ ($s = 1, 2, ..., m$), is a categorical random variable which can take the following values, 00 - if probe $s$ does not hybridize with neither fragment, 10 - if probe $s$ hybridizes with larger fragment only, 01 - if probe $s$ hybridizes with smaller fragment only, 11 - if probe $s$ hybridizes with both fragments, with probabilities depending on $\theta_{ij}$.

## 2 Full Bayesian analysis

In this previous approach, it was assumed that fragment lengths $L_i$ and $M_j$ are constants such that $\sum_i L_i = \sum_j M_j = N$, where $N$ is the genome size. However, in practice, the lengths of the fragments and hence the genome size are subject to experimental errors and hence it can be argued that they should be considered in the analysis as random, rather than as constant. Taking advantage of previous work (Gonçalves et al., 2005) we propose a full Bayesian analysis where fragment lengths are considered as random variables. Instead of the constants $L_i$ and $M_j$, we consider two random variables $X_i$ and $Y_j$ to describe the lengths of fragments $A_i$ and $B_j$. We
assume for $X_i$ a normal distribution with mean $L_i$ and standard deviation equal to $rL_i$ and for $Y_j$ a normal distribution with mean $M_j$ and standard deviation $rM_j$ (where $r$ is the coefficient of variation (CV)). This way of writing the standard deviation is motivated by biological practice. In fact, small fragments can be measured very accurately but large fragments can be associated with larger measurement errors. Values of $r$ between 1 and 5% seem to be appropriated. In physical mapping of *O. oeni*, strain GM, the maximal relative measurement error (or CV) was estimated to be less than 4.9% for fragments ranging from 6 to 1195 kb (Zé-Zé et al., 2000).

Comparing the PSU-1 physical map with sequencing data from the whole-genome (Mills et al., 2005), it can suggest a value of $r$ less than 5%.

We still assume for $\theta_{ij}$ a mixed prior probability distribution, however, $p_{ij}^0$, $p_{ij}^1$ and $p_{ij}^2$ are now random variables since they are functions of the random lengths. The derivation of these probabilities can be found in Gonçalves (2004). Combining this information with hybridization data with a set of $m$ probes, we can obtain posterior probabilities. Relatively to hybridization experiments, we assume that the number of hybridizations of a probe $s$ ($s = 1, 2, ..., m$) along the genome is a Poisson process with rate $\lambda_s$. Accordingly, the probability that a probe $s$ does not hybridize in a region of length $M$ of a genome is given by $q_s = e^{-\lambda_s M}$. We also assume that this rate $\lambda_s$ is random with a diffuse gamma prior. Additionally, we consider that each probe hybridizes independently of all the other probes. Our aim is to compute the posterior probabilities and for that we implemented all these specifications in WinBUGS14 program. After 5000 burn-in iterations to ensure that the Markov-Chain Monte Carlo (MCMC) algorithm had properly moved away from its starting values and had converged, another 6 000 iterations were performed to obtain the posterior distribution of $\theta_{ij}$.

3 Application to the Genome of *O. oeni* strain GM

*Oenococcus oeni* is a lactic acid bacterium occurring naturally in wine and related habitats (Dicks et al., 1995, Zé-Zé et al., 2000). Chromosomal DNA of strain GM was digested with five different enzymes *Asc*I, *I-Ceu*I, *Fse*I, *Not*I e *Sfi*I. These enzymes produced 3, 2, 7, 12 and 18 fragments, respectively. The size of the *O. oeni* GM chromosome was estimated to be 1932 kb. DNA sequences used as probe were allocated, by Southern hybridization, to the restriction fragments generated by the five enzymes.

For more details about data and experiments see Zé-Zé (2000).

Here we present a small example with some pairs of fragments used in the GM map. The enzymes under consideration are *Not*I and *I-Ceu*I and fourteen probes were used.

In our example, the results obtained, with $r = 5\%$, are presented in Table 1. It seems clear that $N_1$ (with length 696) cannot neither in $C_1$ (with length 1206) nor in $C_2$ (with length 726) must be situated either on the...
union $C_1C_2$ or $C_2C_1$, and hence should be partial overlapped with both fragments of I- CeuI. Since $N_2$ (with length 350) is not totally contained in $C_2$ and since the posterior probability of non-overlap with $C_2$ is very high, it seems evident that should be contained in $C_1$ (the posterior probability of total overlap is 0.498). Again $N_3$ (with length 202, like $N_1$, should be over one of the unions of the two fragments of I- CeuI, since it cannot be totally contained either in $C_1$ or in $C_2$.

Other conclusions relatively to the other fragments of NotI can be similarly reached, helping the biologist to solve the combinatorial puzzle of inferring the arrangement of restriction fragments.

Acknowledgments: This work is partially supported by POCTI/MAT/44082/2002

References


Bayesian evaluation of times to diagnosis in women with mammograms indicating suspicion of malignancy

C. Armero, A. López-Quílez and R. López-Sánchez

1 Departament d’Estadística i I.O., Universitat de València, C/ Doctor Moliner, 50, 46100 Burjassot (València), Spain

Abstract: Breast screening through mammographies are the most popular method of discovering breast cancer at a very early stage. Generalized linear mixed models (GLMMs) are used in order to examine the waiting period from a positive mammogram for cancer until the confirmatory diagnosis in hospital. Markov Chain Monte Carlo (MCMC) algorithms are applied for Bayesian estimation by simulating the posterior distribution of the parameters and hyperparameters of the model through the free software WinBUGS.

Keywords: Bayesian statistics; breast screening; generalized linear mixed models; Markov chain Monte Carlo methods.

1 Introduction

Breast cancer is one of the diseases with more impact on health in women in developed countries. The most popular method to discover it at a very early level is breast screening. It lies in a mammogram that is an x-ray of each breast. When a woman has a mammogram indicating a well-founded suspicion of malignancy, the quality of pathological services, providing quick and reliable diagnoses and prognostic information, is essential for the treatment success. Consequently, quantitative studies to evaluate delays in the confirmatory diagnostic processes is desirable to increase the quality and benefits of the service.

The reference framework of this work is the Comunitat Valenciana Breast Cancer Screening Program (CV-BCSP, from now on). The Comunitat Valenciana is an Spanish autonomous region located in the east of Spain, has a population of about 4 203 thousands of inhabitants and has completely taken on the management of all health services. The first step of a woman in the CV-BCSP occurs in the breast screening unit and involves a mammogram of each breast. In the case of an abnormal mammogram, the woman is transferred to the hospital in order to confirm the diagnosis.

In this paper we exclusively concentrate on two confirmatory procedures involving a biopsy: stereotactic core biopsy (CoreB from now on) and wire-guided open biopsy (OpenB from now on). CoreB was developed to provide
a non-surgical means to diagnose breast cancer, whereas OpenB is most aggressive and expensive but provide a more definitive diagnosis.

We analyze the time period between the detection of the abnormal mammography and the definitive diagnosis: people in charge of the CV-BCSP were interested in evaluating and comparing both periods in connection with the qualitative evaluation of the mammography lesion and the hospital where the diagnosis procedure is carried out.

2 The mammogram data

Our data bank comes from the CV-BCSP and records all women with an abnormal mammography all year 2002 round with a final diagnosis as a result of a CoreB or an OpenB. All observations are anonymized and contain, among others, information on the time, in days, required to determine the final diagnosis (time to diagnosis from now on), the mammographic lesion detected and the hospital (25 in all in the Comunitat Valenciana) where the tests are conducted.

The distribution of the time to diagnosis presents a heavy right tail in both confirmatory procedures and a clear concentration of the data on small values. For CoreB times, the median is 23 days with a standard deviation of 48.41 days, and a 25% of the women are waiting for more than 41 days for the confirmation of the diagnosis. In the case of the OpenB times, the median is nearly twice, 41 days, and the data are a bit more concentrated with a standard deviation of 33.22 days, being 62.75 days the 75th percentile.

3 Statistical modelling

Time to diagnosis remind us times to failure in reliability studies, life times in biomedical applications, etc; that is, the so-called times to event data (Crane et al., 2002). After ruling out a logarithmic transformation we have assumed the gamma model as appropriate to describe the stochastic behaviour of the time to diagnostic with each procedure. Bayesian generalized linear mixed models (Zhao et al., 2004) are considered to relate the expected time to diagnostic with the mammographic lesion, a fixed-effect factor, and the hospital, a random-effect one. There are not systematic differences between hospitals, and consequently, they have been considered as a random sample from a common general population.

We have chosen the logarithmic link function basically because it has operated very well with our data. It is a very usual link for the gamma distribution due to the close relationship with the logarithmic transformation of the response variable data. In order to fully manage our statistically modelling we have elicited diffuse prior distributions for the parameters of the model and built a hierarchical model for the random hospital effect.
Different assumptions for the variance of the time to diagnosis has been examined: constant variability (parameterization 1), equal coefficient of variation in all hospitals (parameterization 2), and variance proportional to expected time to diagnostic (parameterization 3). Markov chain Monte Carlo (MCMC) algorithms are applied for estimation by simulating the posterior distribution of the parameters and hyperparameters of the model through the free software WinBUGS (Spiegelhalter et al., 2003). The deviance information criterion (DIC) introduced by Spiegelhalter et al. (2002) has been used for model selection. This criterion weights up the goodness of fit and the complexity of the selected model, and can be considered as a generalization of the Bayesian information criterion (BIC) and Akaike information criterion (AIC). The final fitted model for CoreB times is achieved with parameterization 2, and for OpenB times the better results has been obtained with parameterization 3. (For more details see Armero, Lpez-Qulez and Lpez Snchez, 2006). So, if $T$ represents the time to diagnosis, our fitted model turns out as:

$$ T_{ijk} \sim \text{Ga}(a_{ijk}, b_{ijk}), \text{ with } \mu_{ijk} = a_{ijk}/b_{ijk}, \sigma_{ijk}^2 = a_{ijk}/b_{ijk}^2 $$

$$ \log(\mu_{ijk}) = \alpha + (l_i - 1) + h_j, \ i = 1, 2, \ldots, 6, \ j = 1, \ldots, 25 $$

where $k$ indexes the woman, $j$ the hospital and $i$ the mammographic lesion ($i = 1$ stands for calcifications that we have chosen as the reference category.
in the model because it is the most numerous, \( i = 2 \) for combined/multiples, \( i = 3 \) for nodules and masses, \( i = 4 \) for architectural distortion, \( i = 5 \) for asymmetric parenchyma and \( i = 6 \) for skin changes). The rest of the model depends on the confirmatory procedure, CoreB or OpenB, selected:

1. **CoreB**:
   \[
   \begin{align*}
   a_{ijk} &= 1/\psi^2, \quad b_{ijk} = (\varphi^2 \mu_{ijk})^{-1} \\
   \alpha &\sim \text{N}(0, 100), \quad l_i - l_1 \sim \text{N}(0, 4), \quad i = 2, 3, 4, 5, \quad \varphi \sim \text{Un}(0.1, 2) \\
   h_j &\sim \text{N}(0, \eta^2), \quad j = 1, \ldots, 25, \quad \eta \sim \text{Un}(0.1, 3)
   \end{align*}
   \]

2. **OpenB**:
   \[
   \begin{align*}
   a_{ij} &= \mu_{ij}/\psi^2, \quad b_{ijk} = 1/\psi^2 \\
   \alpha &\sim \text{N}(0, 100), \quad l_i - l_1 \sim \text{N}(0, 4), \quad i = 2, 3, 4, 5, 6, \quad \psi \sim \text{Un}(0.2, 10) \\
   h_j &\sim \text{N}(0, \eta^2), \quad j = 1, \ldots, 25, \quad \eta \sim \text{Un}(0.1, 3)
   \end{align*}
   \]

Figure 1 shows 95% posterior intervals of the expected times to diagnostic for CoreB and OpenB procedures with regard the reference hospitals. Results indicate that longer times are associated to core biopsies with a lot of differences between hospitals. Also, longer waiting times are observed in mammograms with calcifications, as many in core as open biopsies.

**Acknowledgments**: This research is partially supported by the Spanish Ministry of Science and Education, under grant MTM2004-03290. The authors thank the Oficina del Plan del Cáncer of the Generalitat Valenciana for the data used in the analysis.

**References**


Spatial survival modelling and the 2001 UK foot-and-mouth disease epidemic

T.C. Bailey\textsuperscript{1} and T.J. McKinley\textsuperscript{1}

\textsuperscript{1} Mathematical Sciences, Laver Building, University of Exeter, EX4 4QE, UK, T.C.Bailey@exeter.ac.uk

Abstract: The potentially large economic impacts of animal disease epidemics have been highlighted in recent years through outbreaks such as foot-and-mouth disease (FMD) in the UK during 2001. We report work from an ongoing project with the Veterinary Laboratories Agency (VLA), Weybridge, UK which is concerned with use of survival modelling to develop dynamic space-time predictions of survivor and hazard functions for individual farm premises as an animal disease epidemic progresses. Such survival analyses could provide powerful insights into the patterns of infection, and assist in optimising various aspects of the operational response activities, such as targeting of ‘at-risk’ premises. We consider various possible model formulations and illustrate a range of these on data based on real and simulated epidemics.

Keywords: FMD; survival models; cure rate models; mixture models; frailties.

1 Introduction and background

The typical aims of epidemiological analyses of contagious animal disease are twofold: firstly to understand the dynamics of the disease in question, and secondly to apply this knowledge to predict the future path of the epidemic and guide the development of rigorous and effective control policies. The 2001 outbreak of FMD in the UK resulted in the slaughter of almost 10 million livestock at an estimated cost of £8 billion (DEFRA, 2005). FMD can be transmitted in various ways, most commonly through direct (or indirect) contact with infected animals or through dispersion by an environmental factor such as wind (e.g. Ferguson \textit{et al.}, 2001). The density and species of animal are known to affect both the rate/magnitude of viral excretion from infected herds and susceptibility in uninfected herds (e.g. Keeling \textit{et al.}, 2001). However many influential factors remain uncertain, such as the effects of control policies and culling.

Various modelling approaches have been used, predominantly based around SEIR (susceptible-exposed-infected-removed) frameworks, both deterministic and stochastic (e.g. Ferguson \textit{et al.}, 2001; Keeling \textit{et al.}, 2001). Other approaches include modelling spatial spread through the use of cellular
automata (e.g. Doran and Laffan, 2005) or space-time point process techniques (Diggle, 2005; Wilesmith et al., 2003). Survival modelling is an approach that provides a flexible framework from which key quantities such as predicted survival times and ‘transmission potential’ (hazard), can be obtained. Although principally modelling the temporal path of the epidemic, spatial aspects can be incorporated through, for example, a specification of a covariate which is space-time dependent and/or spatially structured random effects (frailties). Survival modelling also provides a tractable method for incorporating censored observations e.g. farm premises that were culled during or remain uninfected at the end of the study period. Censoring takes on additional importance in epidemic models where exposure to the disease changes over space as well as time, since the population ‘at-risk’ is then also dynamic. The inclusion of frailties in such models can be most flexibly handled in a Bayesian framework using Markov Chain Monte Carlo (MCMC) sampling. This also allows full posterior predictive distributions for survival times to be obtained.

2 Model specification

Exploratory work by VLA provided best fitting functional forms for ‘viral excretion rates’ over time associated with different types and sizes of infected animal herds. From these ‘infectivity functions’, the viral load from each infected herd can be estimated, based on size, species and time since infection. Given an infection time \( t_j \) for an infected premise (IP) located at \( s_j \) \((j = 1, \ldots, J)\), a space-time covariate (denoted ‘confirmed infective load’, or CI) can be calculated for any point location \( s \) at any time \( t \) as:

\[
CI(s, t) = \sum_{j|t_j < t} \left[ \sum_k \mathcal{I}_k(t - t_j, n_{jk}) \right] \omega(d_j, \tau),
\]

where \( \mathcal{I}_k \) is the known infectivity function for species \( k \) (e.g. sheep, cattle or pigs), \( n_{jk} \) is the herd size, \( d_j \) is the distance between \( s \) and \( s_j \) and \( \omega \) is a suitable kernel spatial smoothing function with a bandwidth \( \tau \). CI defined in this way has various advantages as a survival model covariate. It is calculated sequentially and as such can adapt to include information from culled premises and new infections (including ‘spark’ infections not obviously attributable to localised spread). It can also be used to focus modelling only on premises whose cumulative exposure to the virus at any time exceeds a pre-determined CI threshold. This implies modelling relative rather than absolute survival time (i.e. time from exposure), and avoids overloading models with large numbers of censored observations relating to premises which are effectively never at risk. Since prediction of the future path of the epidemic is key, a discrete para-
metric form was chosen for the distribution of the survival times so that:

\[ S_i(t) = \exp \left[ -\sum_{m=1}^{t} \left\{ \lambda_i(m-1)(m^\alpha - (m - 1)^\alpha) \right\} \right] , \]  
(1)

where \( i \) denotes premise, \( m \) denotes time point, \( \alpha > 0 \) is a scale parameter and \( \lambda_i(m-1) = \exp(\beta_0 + \beta_1 CI_i(m-1) + \beta_2 AD_i(m-1) + \psi_i + \theta_i) \). Here AD denotes uninfected animal density as a surrogate for susceptibility and \( \psi_i \) and \( \theta_i \) are frailties that are spatially unstructured and structured respectively.

The above model was fitted to data covering recorded infections from the first 43 days of the Devon FMD epidemic using point locations of premises, associated numbers and species of animals, infection or censoring dates and cull status. A Bayesian framework and MCMC was used to derive posterior predictive survival times for the remainder of the epidemic. Although the overall model fit was acceptable, it was found to significantly overpredict survival times for subsequently infected premises. A possible explanation is evidence of a ‘resistance to infection’ being exhibited by some premises, over and above that which is explained by the susceptibility covariate used.

In the absence of any other available susceptibility covariates, it is necessary to investigate other modelling frameworks which incorporate unknown resistance to infection. One formulation is a ‘long-term survivor’ or ‘cure rate’ model, which incorporates an unknown ‘immune’ proportion of premises into the model and fits a survival distribution to the susceptible proportion only (Maller and Zhou, 1996). The model specification used the approach of Banerjee and Carlin (2004), i.e. \( S_i(t) = p + (1-p)SS_i(t) \), where \( p \) is the cure parameter (probability of resistance), \( SS_i(t) \) is the survivor function for the susceptible group of the form (1). Covariates are included as before in the scale parameter \( \lambda \) or in the cure parameter \( p \) if required. Although this removes the potential for resistant premises to ever experience infection, it reduces confounding censored information so potentially resulting in more accurate parameter estimates and better prediction of genuine localised spread. Alternatively, we can extend this idea to allow some chance of infection in the resistant group by forming a two-group mixture model i.e. \( S_i(t) = pSR_i(t) + (1-p)SS_i(t) \), where the additional \( SR_i(t) \) is the survivor function for the resistant group.

To test the relative performance of these different formulations, we constructed a series of simulated epidemics within the framework of the locations of the Devon premises. The epidemics were simulated using one survival process for the susceptible group (based on CI), and a different survival process, representing spark infections, for all premises. Varying levels of resistance were used, based on a simulated biosecurity measure. Models were fitted to the data censored one week from the end of the full period simulated so that predicted and actual infection dates in that week could be compared. Due to the time/infection dependent nature of the CI covariate, predictions had to be derived by using the parameter values and
Table 1 shows some summary results from our investigations. The first two columns compare the percentage of UIPs (uninfected premises) that actually became infected in the week simulated, with the corresponding percentage predicted to become infected in that period by the various models. The third column shows the percentage of the actual infections in the simulated week that were correctly identified. The final column shows the percentage of UIPs correctly specified into susceptible and resistant groups. Overall it can be seen that as the level of resistance increases the conventional model breaks down completely and substantially overpredicts survival time. The other two approaches perform consistently better and although the accuracy of individual premise predictions is still not high, they correctly categorise approximately 80% of premises as either resistant or susceptible. A further desirable feature of the latter two approaches is that in the presence of no resistance they satisfactorily reproduce the conventional model results. Repeated simulations support the patterns identified here. So broadly, the results indicate that cure or mixture modelling approaches are potentially much more useful in the modelling of survival in FMD (and other similar epidemics), as compared with the conventional model.

To illustrate the range of ways in which this modelling strategy can usefully be applied, a mixture model was fitted to a simulated outbreak and various quantities of interest were extracted. The mixture model was chosen on the basis of the above results and because it allows multiple survival processes to be modelled. The simulation used medium levels of resistance and a standard cull policy (24hrs for IPs and 48hrs for any UIP within a 2km radius of an IP). In order to allow the epidemic to get established, no culling

T.C. Bailey and T.J. McKinley 77
was introduced until two weeks after the first IP report date. Again the Devon framework was used as the basis for the simulation, which involved 2000 premises with some spatial structure in the resistance to infection in approximately 50% of these.

We first demonstrate how the model can be used to describe the evolution of the epidemic over the first 35 days. Accordingly the model was fitted at various time points during this period and figures 1 a), b) and c) show resulting hazard maps for days 14, 22 and 30 respectively, formed by spatially smoothing the posterior mean probability of infection in the next day given survival up to that point. A plot of the hazard over time can be formed by fitting the model on a day-by-day basis and averaging the individual premise hazards over space at each time point. Figure 1 d) shows the actual number of new IPs at each day with a scaled version of the hazard over time superimposed. This indicates that the retrospective fit seems to capture the actual course of the epidemic reasonably well.

Predictions for the future path of the epidemic were obtained using the method described earlier. In this case the model was fitted to the simulated epidemic at two weeks and figure 2 a) is a spatially smoothed map of the posterior predictive mean probability of infection in the next week. Alternatively, we may also rank current UIPs based on the order in which they are predicted to become infected. Figure 2 b) is a map of the mean probability of belonging to the next fifteen UIPs that become infected based on the predictive posterior distribution of the rank order of premise infection. Superimposed on both figures are the locations of the next fifteen actual infections and it can be seen that these locations fall within the areas predicted by the model.

In addition to survival times the model can return predictive posterior distributions for the probability of resistance to infection. Figure 3 is a plot of predicted resistance from the model fitted at two weeks using a 75% credible interval to determine whether premises are classified as resistant, susceptible or unsure.

Clearly, a key question that arises from these sorts of predictive results is to what extent surveillance can be successfully targeted during the course of the epidemic. To investigate this, three similar epidemics were simulated - one with no culling, one with the standard ring cull policy described earlier, and a final one in which targeted culling was used based on model predictions. The targeting involved culling IPs and also culling the top fifteen ‘at-risk’ UIPs. The latter was based on the predictive posterior distributions of the ranks from sequential updating of the model at weekly intervals. In both cases where culling is involved, IPs were culled within 24hrs of report and targeted UIPs within 48hrs of report with neither cull policy introduced until two weeks after the first reported infection.

Figure 4 shows the numbers of infections at each day in the epidemic for each of the three cull strategies. The corresponding total numbers of infections and culls are shown in table 2. Although the total number of infections
FIGURE 1. Hazard plots

FIGURE 2. Predicted hazard plots
was higher in the case of the targeted cull the total number of culls was less than half that required by the ring cull to bring the numbers of infections down to the same level.

4 Conclusions

Overall the results presented here suggest that the cure and mixture model approaches offer various advantages over the conventional survival approach when dealing with animal disease epidemic data. In particular they incorporate the notion of resistance to infection into the model and where there is no evidence in the data to support this then they reduce to the conventional model with little loss of accuracy. We have also illustrated how these models may be fitted sequentially to describe the evolution of the epidemic and predict its future course, leading to the possibility of target culling as a more efficient means to reduce the transmission potential of the disease. Further work in progress includes the application and refinement of these approaches using real data from the 2001 UK FMD epidemic.

Acknowledgments: to P. Durr and M. Arnold, VLA, Weybridge, UK.
FIGURE 4. Path of epidemic over time with different cull strategies

References


Doran, R.J. and Laffan, S.W. (2005). Simulating the spatial dynamics of foot and mouth disease outbreaks in feral pigs and livestock in Queensland, Australia, using a susceptible-infected-recovered cellular automata model. *Preventive Veterinary Medicine, 70*, 113-152.


Longitudinal models for shape data

Sarah Barry\textsuperscript{1} and Adrian Bowman\textsuperscript{1}

\textsuperscript{1} Department of Statistics, University of Glasgow, 15 University Gardens, Glasgow, G12 8QW, Scotland, UK

Abstract: We apply a pairwise modelling approach to multivariate longitudinal shape data from a study comparing the facial shapes of children with cleft-lip and palate (post-surgery) to a control group at 6, 12 and 24 months of age. We display estimated group differences for the coordinates of each landmark over time with confidence intervals and use these to study the persistence of differences in each part of the nose profile. We present estimated AIC values constructed by evaluating the full likelihood at the pairwise modelling parameter estimates and use these to compare models, specifically with respect to investigation of group and gender effects.

Keywords: Shape; longitudinal; pairwise.

1 Introduction

1.1 Shape analysis

Shape analysis has been used for many different applications in various fields, including computer science and medicine as well as statistics. Shapes are generally defined by landmarks, which are points of correspondence on objects that match between and within populations. Dryden and Mardia (1998) give a full description of the methods used to standardize and analyse shape data.

The space in which a shape exists is a hypersphere. Tangent space is a linearized version of shape space at a tangent to the point on the hypersphere that usually corresponds to the average shape. If points are assumed to be close together in shape space then the projections onto tangent space can be assumed to approximate the shape information. Euclidean distances in tangent space are therefore good approximations to distances in shape space and general multivariate methods may be applied (Dryden and Mardia, 1998).

1.2 Longitudinal shape data

When shape data are longitudinal, the ideal situation would be to fit a joint mixed effects model to the tangent coordinates to investigate relevant group

82
differences and changes over time. For example for $m$ outcomes, which are defined in this case as individual tangent coordinates, to be modelled jointly we assume that $Y_{ri}$, the $n_{ri}$-vector of repeated measurements across time for subject $i$ for outcome $r$, is distributed as

$$
Y_{ri} \sim f_{ri}(y_{ri}|b_{ri}, \psi_r),
$$

for $b_{ri}$ a vector of random effects and $\psi_r$ a vector of fixed effects for outcome $r$. If the $m$ random effects vectors $b_{ri}$ are stacked into a vector $b_i$ and $D$ represents its covariance matrix, then it is clear that as the number of outcomes increases the dimension of $D$ becomes very large and its parameters difficult or impossible to estimate (Molenberghs and Verbeke, 2006).

Fieuws and Verbeke (2006) suggested instead a pairwise modelling approach, which preserves the advantages of the joint modelling approach while avoiding the computational difficulties inherent in estimating the parameters of $D$ for a large number of outcomes. We will present this approach as described by Fieuws and Verbeke (2006) and adapt it to data obtained from a study of facial shapes of infants with cleft lip and palate (Hood et. al, 2004).

2 Pairwise modelling approach

The Fieuws and Verbeke (2006) pairwise modelling approach involves fitting $m(m - 1)/2$ bivariate models for all possible pairs

$$(Y_1, Y_2), (Y_1, Y_3), \ldots, (Y_1, Y_m), (Y_2, Y_3), \ldots, (Y_2, Y_m), \ldots, (Y_{m-1}, Y_m),$$

where $Y_r$ is a vector of measurements of the $r^{th}$ outcome across individuals and times. If we have $m$ outcomes (here $m$ = the number of landmarks multiplied by the number of dimensions) and refer to a combination of outcomes $(r, s)$ as pair $p$, then the sum of log-likelihoods

$$
\sum_{i=1}^{N} l_{rsi}(Y_{ri}, Y_{si}|\theta_{r,s}) = \sum_{i=1}^{N} l_{pi}(\theta_p)
$$

is maximised for each $p$. If $\theta^*$ is defined as the vector of all the unknown parameters (regression coefficients and random effects’ covariances) in the joint model and $\theta$ as the stacked vector of all $\theta_p$, then it is possible to find a single estimate for each parameter by averaging over all the pair-specific estimates in $\hat{\theta}$ (i.e. the estimates from pairs $(r, s)$ and $(r, s')$). However the standard errors of the averaged estimates may not be found in the same way and, in any case, the pair-specific estimates are correlated and this must be taken into account.

Fitting all of the pairwise models is equivalent to maximising a pseudo-likelihood and under such a framework $\text{Var}(\hat{\theta}) = J^{-1}KJ^{-1}/N$ asymptotically, where $N$ is the number of individuals and $J$ and $K$ are matrices of
second and first derivatives, respectively, both averaged across individuals. If $A$ is a matrix of appropriate coefficients used to calculate the average estimates, then the required vector of parameter estimates $\hat{\theta}^* = A\hat{\theta}$ and, approximately, $\text{Var}(\hat{\theta}^*) = AJ^{-1}KJ^{-1}A'/N$.

This approach is described in detail in Fieuws and Verbeke (2006), who showed, using a simulation study, that the efficiency of the pairwise approach is close to that of the fully joint model.

3 Application to cleft lip and palate data

3.1 Introduction

We have data from a longitudinal study on infants with cleft lip and palate, which captured 3-dimensional images of 49 children with unilateral clefts and 100 age-matched non-cleft controls at age 3 months (prior to primary repair), and 6, 12 and 24 months using a sophisticated stereophotogrammetry system (Ayoub et al., 2003). The background of the study is given in Hood et al. (2004). An example of a low resolution surface mesh of points describing an individual face is displayed in Figure 1. Landmarks are placed at anatomically meaningful points (such as around the lips and the base of the nose) on a high resolution surface mesh and it is these that are used to describe the shape in an analysis.

![Low resolution surface mesh of an infant from side and front views](image)

There is interest in the differences between the cleft and control groups and in how the shapes change over time as the children grow, particularly after the cleft group undergoes surgery. Interest also lies in whether there are shape differences between males and females, both initially and over time.

3.2 Applying the pairwise models

The initial work has been done on 4 landmarks in 2 dimensions and for time points 6, 12 and 24 months (since we wish to study how the groups change...
over time after the cleft group undergoes surgery). The chosen landmarks describe the profile of the nose (hence the $x$-coordinates are not necessary) and are all placed on the midline of the face. They are displayed in profile view in Figure 2. The codings represent the names of the landmarks: $ls$ is the top of the upper lip, $sn$ the base of the nose, $prn$ the tip of the nose and $n$ the midpoint between the eyes.

We applied the pairwise model approach to our data using a program written in R (R Development Core Team, 2004). This calculates the relevant derivatives in the $J$ and $K$ matrices in $\text{Var}(\hat{\theta})$ using the deriv function in the stats package. If $Y_{ij}(t_j)$ is the value of a single coordinate for individual $i$ at time $j$ months after the beginning of the study and time is modelled as a factor, then the model is given by

$$Y_{ij}(t_j) = \beta_0 + b_{0i} + \beta_1\text{group} + \beta_3t_{12} + \beta_4t_{24} + \beta_5\text{group} \times t_{12} + \beta_6\text{group} \times t_{24} + \epsilon_{ij}(t_j),$$

where $t_{12}$ and $t_{24}$ are indicator variables representing the 12 and 24 month time points respectively and $b_{0i}$ is a random intercept. The random intercepts, $b_{0i}$, follow a Normal distribution with mean 0 and some variance $\sigma_i^2$ where the $\sigma_i$ values are allowed to be different for each coordinate. The intercepts are also allowed to be correlated between coordinates within individuals.

The results of particular interest are the differences between the cleft and control groups over time. These are displayed for the individual coordinates in Figure 3. In the vertical, or $y$-direction ($ls_y$, $sn_y$, $prn_y$, $n_y$), there are significant differences between the groups for all 4 landmarks at 6 months.
In the horizontal, or $z$-direction (i.e. coming out of the face, describing the shape of the profile), there are no significant differences between the groups for any of the landmarks at this time point.

As time progresses the groups come closer together in the vertical direction but by 24 months of age there are still significant differences between them for 3 of the landmarks ($ls$, $sn$ and $n$) in the $y$-direction and 2 of the landmarks ($ls$ and $sn$) in the $z$-direction (although the difference for $ls$ is marginally significant). The magnitude of the differences is smaller in most cases at 24 months than at 6 months. The 12 month group differences are very similar to those at 24 months and it is likely that a model with group means for 6 months and post 6 months (i.e. for all 12 and 24 month responses) would be suitable.

Another way of presenting the results is to plot the estimated mean shapes with 2-dimensional 95% confidence regions for the cleft and control groups at each time point, approximating the transformation back to shape space by adding on the empirical mean shape. These are displayed in Figure 4 for each of the time points. The tight confidence regions around the control group estimates in particular reflect the high precision afforded by a large sample size and that, relative to the overall scale of a face, there is very little variation in a particular landmark between individuals.

These plots show consistent differences between the groups over time in the $sn$ and $ls$ landmarks, in other words the points at the base of the nose and the top of the upper lip. The other two landmarks at the tip of the nose ($prn$) and between the eyes ($n$) do not show such persistent differences. These results corroborate those obtained from Figure 3. In terms of
facial shape the general difference between the groups seems to be that the cleft children have significantly shorter upper lip height, as measured from the base of the nose to the top of the upper lip. It appears that despite surgery there remain differences between the groups in the lower part of the profile that persist over time, but that they are somewhat reduced as time progresses.

![Figure 4](image_url)

**FIGURE 4.** Two-dimensional 95% confidence regions for the cleft and control group means at each time point. The shaded confidence intervals correspond to the cleft group and the unshaded (and generally smaller) to the controls. Where the control confidence regions are not obviously visible, they lie on the boundary of the corresponding cleft region and are very tight.

### 3.3 Comparing models

Since the results using this method are obtained by fitting individual models for each pair of outcomes \( (r, s) \) and summarising the results across all the outcomes, the log likelihood of the fully joint model evaluated using the parameter estimates obtained from the pairwise models should give an estimate of the log likelihood for the overall fit. This can then be used to calculate the AIC for a particular model and subsequently to compare models.

Some interest lies in whether there are differences between males and females in facial shape and as such, models may be fitted with gender effects.
TABLE 1. *AIC values for models with different gender effects (plus the effects in (1).*

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Gender + gender:t12 + gender:t24 + gender:group</td>
<td>-15732</td>
</tr>
<tr>
<td>(B) Gender + gender:t12 + gender:t24</td>
<td>-15736</td>
</tr>
<tr>
<td>(C) Gender + gender:group</td>
<td>-15747</td>
</tr>
<tr>
<td>(D) Gender</td>
<td>-15755</td>
</tr>
<tr>
<td>(E) No gender effects</td>
<td>-15755</td>
</tr>
</tbody>
</table>

Graphs similar to those in Figures 3 and 4 show no evidence of an effect of gender on average, over time or that differs between groups (not shown). Table 1 gives the AIC values for models containing various gender effects, in addition to the main and interaction effects of group and time in the model in (1). Comparing models B and C against A shows that neither the gender and time interactions nor the gender and group interaction show evidence of being important effects in addition to the other. Comparison of model D with models B and C shows that a model with a main effect of gender only is more appropriate than either of the models with gender interactions; and finally comparing models D and E indicates that there is no evidence of a gender effect at all.

It may be of interest to note that comparing the estimated AIC of model E, which corresponds to the model in (1) above, with the AIC of such model following removal of the group and time interaction terms (-15666) indicates a considerably better fit for model E, which would be expected from the results in Figures 3 and 4. This is therefore further evidence that the group:time interaction terms represent important effects in the model.

4 Conclusion

The pairwise modelling approach allows large numbers of outcomes to be jointly modelled, without the computational difficulties of fitting a fully joint model. This is extremely useful as it allows analysis of the shape as a whole, rather than aggregating the landmark information into an overall score to capture particular features such as asymmetry. Differences between groups and over time may be estimated and plotted back in shape space to give a visual interpretation. We can also estimate AIC for the overall model which can be used to compare with other models.

Our future work will involve applying this model to landmarks in 3 dimensions and to higher numbers of landmarks. We will also apply it to shapes described by curves, which provide richer information than landmarks.
References


A measure of progress in air quality management: air health indicator

Sabit Cakmak\textsuperscript{1}, Wanchao Li\textsuperscript{1}, Claudia Blanco\textsuperscript{2}, Marcelo Farnandez\textsuperscript{2} and Rick Burnett\textsuperscript{1}

\textsuperscript{1} Health Canada, 137 EHC, Tunneys Pasture A.L 0801A, Ottawa, ON. Canada, sabit.cakmak@hc-sc.gc.ca
\textsuperscript{2} Conama, Metropolitana De Chile, Santiago, Chile

Abstract: Although progress has been made in the last few decades at reducing ambient concentrations of air pollutants, scientific evidence suggests that there remains a risk to human health from exposure to these pollutants. In this paper, we propose a health indicator, which can be used as a measure of progress in air quality management over time. The Air Health Indicator (AHI) is defined as the percentage of the number of daily deaths attributable to exposure to the pollutant of interest. The AHI is proportional to the level of risk, estimated using a generalized additive Poisson regression model with statistical smoothing techniques, and the level of the pollutant of interest. The model is adjusted for time and location-specific covariates such as meteorological conditions (temperature, humidity, etc), and day of week. The AHI is illustrated with an example of how changes in PM10 levels in 7 Chilean communities from 1997 to 2003 have translated into changes in the percent of non-accidental deaths due to this pollutant over time.

Keywords: Air pollution; mortality; air health indicator.

1 Introduction

Historical extreme air pollution events such as those experienced in London in the 1950s and 60s clearly demonstrated the potential of ambient air pollution to cause exacerbation of cardio-respiratory disease, manifested as pre-mature mortality and admission to hospital. In the intervening years, considerable efforts have been made to reduce pollution from the combustion of fossil fuels and industrial activity. Although these pollution mitigation strategies have been largely viewed as successful, recent evidence from population health studies continues to identify ambient air quality as a population health concern (1). Progress in pollution mitigation can be measured in several ways. Primary pollutant emissions into the atmosphere can be measured or estimated and tracked over time and ambient air concentrations of primary and secondary pollutants can be measured at representative sites. It is desirable, however, to be able to demonstrate that
societal expenditures directed to improve air quality have been effective, by yielding improvements in population health as a consequence of pollution mitigation measures. Our objectives is to determine to what extent air pollution, PM10 in this study, is associated with non-accidental deaths in Chile. The extent will be introduced as an Air Health Indicator (AHI) for PM10.

2 Statistical Model for Risk of Air Pollution on Public Health

Our main goal is to assess the risk of air pollution concentrations on population health over time. Our base model includes time, temperature, day of the week, and air pollution level as risk factors. Air pollution related mortality is believed to depend on time and weather in non-linear fashion. A generalized additive Poisson regression model can be used to relate these two time series. It is assumed that the observations are independent and that the logarithm of the expected number of daily deaths is modeled by an additive function of the independent variables. The outcome is assumed to depend on a sum of non-parametric smooth functions for each variable. Consider a time series of the daily number of death on day t within community i in year j, \( Y_{ij}(t) \), and corresponding concentration of an ambient air pollutant, \( X_{ij}(t) \). These two time series can be related through the following model

\[
\log \left\{ E\{y_{ij}(t)\} \right\} = g_{1ij}(t; \delta) + g_{2ij}(w(t); \phi) + \beta_{ij}x_{ij}(t),
\]

where \( g_{1ij} \) and \( g_{2ij} \) are non-linear smoothing functions and describe the potentially non-linear association between time and weather variables, and deaths, respectively. These functions are known up to the parameters \( \delta \) and \( \phi \), which are to be estimated from the data. The unknown parameter \( \beta_{ij} \) represents the logarithm of the relative rate of death attributable to a unit change in air pollution specifically for community (or location) i in year j. A time dependent estimate of the association between pollution exposure and mortality, \( \beta_{ij} \), is required for our Air Health Indicator. The model specification is completed by assuming that the variance of \( Y(t) \) is \( \pi E[Y(t)] \), where \( \pi \) is the amount of over (under) dispersion relative to Poisson variation (\( \pi = 1 \)). The unknown parameter \( \beta \) is estimated by iteratively re-weighted least squares methods (2). Our interest is in describing how \( \beta(t) \) varies smoothly in time. We expect a set of n risk estimates to provide information on a time trend in risk in that selected location.

3 A Measure of Progress in Air Quality Management: Air Health Indicator

We consider that we have made progress in the management of air quality if the population health impacts of atmospheric pollution have been reduced.
The Air Health Indicator (AHI) is defined as the percentage of daily deaths attributable to exposure to the pollutant being modeled. The AHI for year $j$ and location $i$ is defined algebraically by

$$AHI_i(j) = \hat{r}_i(j) \hat{X}_i(j) \times 100,$$

where $\hat{r}_i(j)$ and $\hat{X}_i(j)$ are a temporally smoothed estimate of risk and temporarily smoothed estimate of the air pollution concentration through the year $j$ respectively. We assume that the risk of air pollution changes smoothly over time. However, some unpredicted factors or estimation error can lead to significant variations year by year. Thus an adjustment of the risk estimates are needed. Suppose we decide to use risk estimates from the past $n$ year to adjust the risk estimate in year $j$. For community specific AHI, to say city $i$, the estimate of the adjusted risk, $\hat{r}_i(j)$ will be

$$\hat{r}_i(j) = \begin{cases} \sum_{m=j-n+1}^{j} w_{i,m} \beta_i(m) & \text{if } j \geq n \\ \beta_i(m) & \text{if } j < n \end{cases}$$

where $w_{i,m}$ is the weight and $\beta_i(m)$ is the risk estimate for that city $i$ in year $m$, respectively. The weight are chosen such that more weight are assigned to those years which are closer to the chosen year $j$, and in this study we choose the tri-cube function for these weights. Uncertainty in the estimates of AHI can be obtained by combining the uncertainties in $\hat{w}_i(t)$, $\hat{r}_i(t)$, and $\hat{X}_i(t)$ using Monte Carlo methods (3). The Air Health Indicator for all locations combined is defined as the weighted sum of the AHI for each location.

4 An Illustration of the Air Health Indicator

The daily numbers of non-accidental deaths in 7 Chilean communities (Les Condes, Cerrillos, El Bosque, Florida, Independentia, Parque, and Pudahuel) were obtained from January 1, 1997 to December 31, 2003 in addition to the daily average values of particulate matter concentrations ($PM_{10}$). The air health indicator is displayed for each community in Figure 1. Fairly steady progress in the management of particulate matter ($PM_{10}$) can be observed for all the communities. The value of the AHI at the beginning of the study period, January 1, 1997, was 8.2% based on a weighted average of the city-specific values, while the value at the end of the study period, December 31, 2003, was 6.5%. Thus we predict that changes in ambient concentrations of ($PM_{10}$) in 7 of Chilean communities contributed to a 1.7% reduction in the percentage of deaths attributed to pollution exposure. We conclude that some progress have been made in the improvement of population health due to improvements in ambient concentrations of particulate matter in the 7 communities examined.
5 Discussion

The AHI could help assess if the most important components of the atmospheric mix were appropriately targeted for reduction. For example, particulate matter is comprised of several physical and chemical characteristics with varying toxicity. Mitigation strategies that affect a single or a limited number of pollutants may not yield expected proportionate improvements in population health if those mass reductions do not coincide with reductions in the most toxic components of particle phase pollution. The AHI can be interpreted as a measure of accountability with respect to air pollution mitigation. Particulate matter ($PM_{10}$) attributable-risk of death declined over time in conjunction with declining ambient concentrations suggesting that these improvements in air quality have in fact translated into improvements in population health.

References


Analyzing sequential spells via conditional survival models

M. Luz Calle\textsuperscript{1}, Anna Espinal\textsuperscript{2} and Guadalupe Gómez\textsuperscript{3}

\textsuperscript{1} Universitat de Vic, Spain
\textsuperscript{2} Servei d’Estadística, Universitat Autònoma de Barcelona, campus Bellaterra, Edifici D, 08193, Cerdanyola del Vallès, Spain
\textsuperscript{3} Universitat Politècnica de Barcelona, Spain

Abstract: Marginal approach is used for the analysis of a sequence of durations times. Multivariate survival times come from a treatment interrupted scheme followed by a group of HIV-infected patients. That is, each individual defines its own sequence of events from the duration of periods with and without treatment. All of them start without treatment and remain in such state until certain conditions fail. In that moment patient starts a new spell with treatment until he/she achieve again the conditions to be free of treatment thus starts the second period free of treatment, and so on. Of special clinical interest is the analysis of the duration of the first four spells and characterize them using a set of covariates. To this end, we propose using the regression method proposed by Prentice, Williams & Peterson (1981) which it is a conditional approach and allows to estimate the effect of a set of baseline covariates on each spell of the sequence.

Keywords: Multivariate failure times; sequential events; conditional regression model.

1 Introduction

Multivariate survival data arise when each subject may experience multiple events or failures. These events may be ordered or unordered. Furthermore, the events may be of the same type, such as, repeated myocardial infarction attacks, and of different types, like infection with HIV and subsequent AIDS diagnosis.

The analysis of multivariate survival data requires different methodologies depending on whether the main goal is the study of the dependence among the different survival times or if the scientific interest centers on the effect of covariates on the risk of failure.

The effect of covariates can be analyzed through frailty models where the dependence among different survival times corresponding to the same subject is specified via random effects (Clayton, 1978). In some situations this frailty component is not necessary because the dependence among survival times is captured by the covariates, that is, the survival times are conditionally independent given covariates. The regression methods based on
this working independent assumption include the marginal approach due to Wei, Lin & Weissfeld (1989), the multiplicative intensity model by Andersen and Gill (1982) and the conditional approach by Prentice, Williams & Peterson (1981), referred hereafter as WLW, AG and PWP, respectively. All these methods are sometimes called marginal approaches because they model the marginal distribution of each failure time variable and the regression parameters are estimated ignoring the dependence between related failure times. However, we will use the term "marginal approach" only when referring to WLW. 

The main difference among these methodologies relies on the definition of the risk sets. While in PWP and AG the risks sets for the \((k + 1)\)th failure are restricted to the individuals who have experienced the first \(k\) recurrences and the times censored before the \(k\)th recurrences are excluded from the risks sets of the \((k + 1)\)th recurrences, WLW follows a competing risk approach where all individuals are at risk of any recurrence during their follow-up.

In this work we are concerned with the analysis of ordered events of two different types where the main goal is the effect of covariates on the risk of failure. We encountered this situation when analyzing data from an AIDS clinical trial known as TIBET (guided-Treatment Interruption BEnefiT) project. Tibet is an ongoing trial in which a controlled intermittent therapeutic strategy aiming to improve the quality of life of HIV-infected patients, as well as to reduce the toxic effects of HAART (Highly Active Antiretroviral Therapy), has been assigned to each patient. This strategy defines a sequence of successive spells on which the patients are alternatively without treatment (state OFF) or under treatment (state ON). The process for an individual consists of a series of sojourns in states OFF and ON and various lifetime variables can be defined within this process.

Of special clinical interest is the analysis of the duration of these spells and characterize them using a set of covariates. To this end, in addition to the spells OFF and ON for the patients of the TIBET trial, there were also collected some baseline and pretherapy covariates.

2 Data

The sequence of spells defined from the TIBET trial may be denoted by \(T_1, T_2, T_3, T_4, T_5, \ldots\). Thus, \(T_1\) is the first spell OFF, that is the time (in weeks) from randomization (and therefore interruption of HAART therapy) to first reinitiation of treatment. \(T_2\) is the length of the first spell ON, that is the time from the first reinitiation until the next interruption. Therefore, \(\{T_{2j-1}, j = 1, 2, \ldots\}\) are all the spells OFF being \(T_{2j-1}\) the time from the \((j)\)th interruption to the \((j)\)th reinitiation. On the other hand, \(\{T_{2j}, j = 1, 2, \ldots\}\) are all the spell ON being \(T_{2j}\) the time from the \((j)\)th reinitiation to the \((j + 1)\)th interruption. Figure 1 represent data for four selected patients.
With respect to the covariate, there are available baseline covariates as well as pretherapy covariates (some HIV-infected patients started the treatment period before the HAART therapy). The set of covariates used in the final models are: A dummy variable of no therapy before HAART (NAIVE, 1=yes); indicator of younger than 35 years (AGE35, 1=yes); indicator of lowest available cd4 count under 200 (CD4N200, 1=yes); the indicator of preHAART viral load under $10^5$ copies/ml (L10VLPH5, 1=yes, taken in log10); and reinitiation cause (REINT2=1 if the cause of starting HAART again is due to a low count of CD4).

3 Multivariate regression survival models

Data arising from the TIBET study may be modelled by multivariate survival time methods. The methods considered in this paper are based on a likelihood function which is built taking into account a proportional hazards assumption for the marginal distributions and a working independence assumption concerning the failure times. Under these assumptions, inferences can be performed using standard univariate methods to estimate the corresponding regression parameters. However, since the consecutive times are not necessarily independent, it is necessary a specific variance estima-
tion correction.

Let $E_1, \ldots, E_m$ be the different consecutive potential events for each subject. There are two possible time scales for the analysis of ordered events: Chronological times and gap times.

Denote by $U_1, \ldots, U_m$, such that $U_1 < \ldots < U_m$, the chronological or total times of the $m$ consecutive events. The times are measured from the same origin, for instance, from the beginning of the study. Let $T_1, \ldots, T_m$, such that $T_1 = U_j - U_{j-1}, j = 1, \ldots, m$, be the gap times or spells between consecutive events.

As in most clinical trials the survival data is subject to right censoring. We will assume that closing the study is the only cause of censoring, and thus independent of all other survival and covariate information. Denote by $C$ the elapsed time from randomization to closing the study.

Chronological times are subject to right censoring if $U_j > C$. Censoring indicators are defined as $\delta_j = 1\{U_j \leq C\}$, for $j = 1, \ldots, m$. Note that $\delta_j = 1$ implies $\delta_1 = \cdots = \delta_{j-1} = 1$. Denote by $\tilde{U}_j = \min\{U_j, C\}, j = 1, \ldots, m$.

Gap times are subject to a more complicated censoring scheme referred as dependent censoring: Gap time $T_j$ is censored if $T_j > C - (T_1 + \ldots + T_{j-1})$. In addition, gap times also have a missing data problem: When gap time $T_j$ is censored, the following gap times, $T_{j+1}, \ldots, T_m$, are missing. We define a missing data indicator, $R_j$, as $R_j = 1$ if $T_j$ is observed or censored and $R_j = 0$ when $T_j$ is missing. In this last case, we set, for notational convenience, $\tilde{T}_j = \tilde{T}_{j-1}$ and $\delta_j = 0$, though these cases will make no contribution to the estimations.

Let $Z_{ji}(t)$ be a vector of covariates for the $i$th subject at time $t$. Suppose that there are $n$ subjects. The observable data for subject $i$ when using chronological times is given by

$$(\tilde{U}_{1i}, \delta_{1i}, Z_{1i}, \ldots, \tilde{U}_{mi}, \delta_{mi}, Z_{mi}), \text{ for } i \in \{1, \ldots, n\}$$

and when using gap times by

$$(\tilde{T}_{1i}, \delta_{1i}, Z_{1i}, R_{1i}, \ldots, \tilde{T}_{mi}, \delta_{mi}, Z_{mi}, R_{mi}), \text{ for } i \in \{1, \ldots, n\}.$$ 

3.1 Working independence assumption

Marginal distributions for each specific failure time $U_j$ are assumed to follow a proportional hazards model,

$$\lambda_{ji}(t; Z_{ji}(t)) \lambda_0(t) \exp\{\beta_j Z_{ji}(t)\}, j = 1, \ldots, m. \quad (1)$$

Under the working independence assumption, that is, assuming that the consecutive times $U_1, \ldots, U_m$ are independent and an independent censoring, the overall partial likelihood function can be factorized as the product
of the marginal partial likelihoods, that is

\[ L(\beta) = \prod_{j=1}^{m} \prod_{i=1}^{n} \left\{ \exp\left( \beta_j Z_{ji}(\tilde{u}_{ji}) \right) \delta_{ji} \right\} \left\{ \sum_{k=1}^{n} H_{jk}(\tilde{u}_{ji}) \exp\left( \beta_j Z_{jk}(\tilde{u}_{ji}) \right) \right\} \delta_{ji} \]  

where \( H_{ki}(t) \) is the risk-set indicator of individual \( i \) for the \( k^{th} \) failure.

When \( U_1, \ldots, U_m \) are not independent given covariates, expression (2) does not correspond to the correct partial likelihood function and is referred to as a pseudolikelihood function. However, if the marginal models are correctly specified, the estimator \( \hat{\beta} \) which satisfies the pseudolikelihood score equation, \( U(\beta) = 0 \), is a consistent estimator of the parameter \( \beta \) (Lin, 1994).

The most common approaches following this working independence assumption are the marginal approach due to Wei, Lin & Weissfeld (WLW), the independent increment method by Andersen and Gill (AG) and the conditional approach by Prentice, Williams & Peterson (PWP). The three methods base the inferences on the pseudolikelihood function (2) and they differ in the way the risk-set indicators are defined, in the use of the individual’s history data and in the use of different time scales.

Wei, Lin and Weissfield approach: This approach is similar to the competing risks approach because treats the ordered event as if they were unordered. That is, it is assumed that an individual is at risk of event \( j \) during its follow-up irrespectively that the event \( j-1 \) has occurred or not. However WLW follows a competing risk approach. The risk-set indicator of individual \( k \) for the \( j^{th} \) failure time is defined as

\[ H_{jk}(t) = I(\tilde{U}_{jk} \geq t). \]

WLW method does not use any information on the individual’s history data and is only appropriate for analyzing total times, that is, times measured from the same origin.

Prentice, Williams and Peterson approach: The method proposed by Prentice, Williams and Peterson is called the conditional approach because it uses information on the individual’s history. It is based on a stratified Proportional Hazards model (Cox, 1972) where the hazard at time \( t \) is assumed to depend on the covariate process up to time \( t \) and, the counting process defined by the number of failures prior to time \( t \). The regression parameters are estimated by maximizing the pseudolikelihood function (2), as in WLW method, but now, the covariate process, \( Z_{ki}(t) \), may contain information on the past individual’s history. In particular, appropriate time dependent covariates may be used to capture the dependence between the sequence of survival times. Also the risk-set indicator contains information on the individual’s past history since it is assumed that an individual cannot be at risk of event \( j \) until event \( j-1 \) has occurred. In this case, the risk-set indicator of individual \( k \) for the \( j^{th} \) failure time is defined as

\[ H_{jk}(t) = I(\tilde{U}_{j-1,k} < t \leq \tilde{U}_{jk}). \]
This approach also allows to analyze the effect of covariates on the gap times between consecutive events. For this time scale, the hazard function (1) takes the form

$$\lambda_{ji}(t; Z_{ji}(t)) = \lambda_0(t - t_{k-1,i}) \exp\{\beta_j Z_{ji}(t)\}, \quad j = 1, \ldots, m.$$  

(3)

We define by $T_{ji} = U_{ji} - U_{j-1,i}$ the elapsed or gap time between the occurrence of event $j-1$ and event $j$. Gap time $T_{ji}$ is right-censored when $U_{ji}$ is and the next gap times, $T_{ki}, k > j$, are missing. The corresponding pseudolikelihood function for analyzing gap times can be written as

$$L(\beta) = \prod_{j=1}^{m} \prod_{i=1}^{n} \left\{ \frac{\exp\{\beta_j Z_{ji}(\tilde{u}_{ji})\}}{\sum_{k=1}^{n} H^*_j(t_{jk}) \exp\{\beta_j Z_{jk}(\tilde{u}_{j-1,k} + t_{ji})\}} \right\} \delta_{ji} R_{ji}$$  

(4)

where the risk-set indicator of individual $k$ for the $j^{th}$ failure time is defined as

$$H^*_j(t) = I(\tilde{T}_{jk} \geq t).$$

**Andersen and Gill approach:** The approach by Andersen and Gill is appropriate when one is interested in the overall effect of covariates for recurrences of the same nature. AG can be seen as a particular case of PWP where the marginal distributions are assumed to have identical baseline hazard functions:

$$\lambda_{ji}(t; Z_{ji}(t)) = \lambda_0(t) \exp\{\beta Z_{ji}(t)\}, \quad j = 1, \ldots, m.$$  

(5)

In this case the overall pseudolikelihood function is given by:

$$L(\beta) = \prod_{j=1}^{m} \prod_{i=1}^{n} \left\{ \frac{\exp\{\beta Z_{ji}(\tilde{u}_{ji})\}}{\sum_{k=1}^{n} \sum_{l=1}^{m} H_{lk}(\tilde{u}_{j-1,k}) \exp\{\beta Z_{lk}(\tilde{u}_{ji})\}} \right\} \delta_{ji}$$  

(6)

where, as in PWP, the risk-set indicator of individual $k$ for the $j^{th}$ failure time is defined as

$$H_{jk}(t) = I(\tilde{U}_{j-1,k} < t \leq \tilde{U}_{jk}).$$

A gap time version of AG model is also possible following similar steps as in PWP model.

### 4 Results

To investigate the effect of the covariates on the consecutive spells we follow the conditional PWP approach where the response vector is the vector of the first four consecutive spells.

In the following there are two tables for the estimates of the covariates for the first two times, that is $OFF_1$ and $ON_1$. Included covariates are NAIVE, AGE35, L10VLPH5, CD4N200 and REINT2.
Since age and being naive are known to be highly correlated we decided to stratify the models by NAIVE. Thus we fit two different models, one for naive patients (NAIVE=1) and one for non naive (NAIVE=0). Table 1 presents the results of the covariate effects on $T_1$, the first time OFF. In the naive group no covariate shows a significant effect on the first time OFF, though age and viral load prehaart are at the limit ($p = 0.07$ and $p = 0.06$, respectively). For the non naive patients a lower viral load level and a higher cd4 nadir are associated with a longer first time OFF.

**TABLE 1. First time OFF.**

<table>
<thead>
<tr>
<th>NAIVE = 1</th>
<th>$\beta_1$ (robust s.e.)</th>
<th>$\exp(\beta_1)$</th>
<th>$P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE</td>
<td>0.717 (0.406)</td>
<td>2.049</td>
<td>0.077</td>
</tr>
<tr>
<td>L10VLPH5</td>
<td>0.980 (0.522)</td>
<td>2.666</td>
<td>0.060</td>
</tr>
<tr>
<td>CD4N</td>
<td>0.043 (0.689)</td>
<td>1.044</td>
<td>0.950</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NAIVE = 0</th>
<th>$\beta_1$ (robust s.e.)</th>
<th>$\exp(\beta_1)$</th>
<th>$P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE</td>
<td>0.363 (0.351)</td>
<td>1.438</td>
<td>0.300</td>
</tr>
<tr>
<td>L10VLPH5</td>
<td>1.119 (0.377)</td>
<td>3.063</td>
<td>0.003</td>
</tr>
<tr>
<td>CD4N</td>
<td>-0.876 (0.289)</td>
<td>0.417</td>
<td>0.002</td>
</tr>
</tbody>
</table>

The covariate effects on $T_2$, the first time ON, are reported in Table 2. Both, in the naive and non naive groups, viral load prehaart and reinitiation cause are associated with the first time ON. In the non naive group, those
patients with lower viral load pre-ART levels and those who reinitiate the medication because of their CD4 counts are associated with longer times ON. In the naïve group, higher viral load pre-ART levels and those who reinitiate the medication because of their CD4 counts are associated with longer times ON. Moreover, the CD4 nadir has a significant effect on $T_2$. Those patients with higher CD4 nadir levels are associated with longer times ON.

5 Conclusions

In this paper, we present an analysis of a multivariate survival data coming from a controlled treatment interruption for HIV-infected patients. Three methodologies are discussed for modelling the effect of baseline covariates on the sequence of times, and the most suitable is used to analyze the data. Main advantage of these methods is in estimating the effect of covariates differently for each spell of the sequence. In our data, for instance, the set of significant covariates for the first time, $OFF_1$ is not the same than for the second time, $ON_1$.

Acknowledgments: Special thanks to IrsiCaixa, HUGTIP where the data come from. Also we are grateful to the GRASS group for the interesting and valuable comments during this analysis.

References


Volatility control on the tails of returns

Joan del Castillo\textsuperscript{1} and Jalila Daoudi\textsuperscript{1}

\textsuperscript{1} Department of Mathematics, Universitat Autonoma de Barcelona, Facultat de Ciencies, Edifici C, Cerdanyola del Valles, 08193 Barcelona, Spain

Abstract: Through the coefficient of variation, we develop a new methodological approach for examining tail behavior using only the existence of two moments. This approach is therefore more stable and more widely applicable than those using kurtosis.

Keywords: Heavy-tailed distributions; coefficient of variation; Laplace and Pareto distributions.

1 Introduction

The most basic model for the logarithm of an asset price of is a Gaussian random walk. This is built by summing independent and identically distributed normal variables (returns). However, from an empirical point of view this assumption is far from perfect. This problem has relevant economic consequences in risk management, in asset allocation and in pricing derivatives.

In this paper we explain non-normality from a point of view that is distinct to those considered to date. Through the coefficient of variation, we develop a new methodological approach for examining tail behavior using only the existence of two moments. We show that our model is reasonable from an economical point of view, because it can be interpreted in terms of the spreads between bidding and asking prices. The model can be used for calculating VaR subject to updating schemes similar to GARCH and for working with several assets in a multivariate setting, as in Hull and White (1998).

2 The problem

Time-series models of changing variance and covariance, called volatility models, are used in finance to improve the fitting of returns on a portfolio with normal assumption. Lévy models for volatility processes have been introduced for analyzing returns. In particular, the normal inverse Gaussian (NIG) model introduced by Barndorff-Nielsen, and the generalized hyperbolic (GH) Lévy processes introduced by Barndorff-Nielsen and Shephard (2001).
In order to fit the returns, let us first consider the symmetric GH-model given by

\[
\begin{align*}
  y_t &= \mu + \sigma_t \varepsilon_t \\
  \log (\sigma_t^2) &= \beta + b_t
\end{align*}
\] (1)

where \(\mu\) and \(\beta\) are constants, \(\varepsilon_t\) is \(N(0,1)\) distributed, and \(u_t = \exp b_t\) is a random effect (independent of scale) distributed as a generalized inverse Gaussian distribution independent of \(\varepsilon_t\). In this case the marginal distribution of \(y_t\) is GH and the particular case of NIG corresponds to inverse Gaussian random effects \((\lambda = -0.5)\).

These volatility models are hierarchical models having random effects on the dispersion structure, as introduced by Lee and Nelder (2006). Therefore, the associated h-likelihood methods can be used to estimate their parameters, see Castillo and Lee (2006).

Figure 1 shows the fitted-probability density functions for exchange rate returns between US dollar and euro-zone currencies for GH, NIG and normal distributions. The data set consists of daily exchange rates from January 1, 1999 (the start date for the euro) to December 31, 2002.

As we can see, there is considerable difference between GH and NIG densities, and both improve the normal approach. Moreover, there is a high spike on zero, which is clearly far from normal. There are rather more outcomes closer to the central value than are predicted by the normality assumption. The NIG distribution in Figure 1 has the same kurtosis as the data set. Therefore, the kurtosis does not correctly explain exchange-rate returns. One of the aims of this paper is to explain this behavior from an economical point of view. We show that it can be interpreted in terms of the spreads between bidding and asking prices. The interpretation given also shows that a model can be constructed with two continuous time processes: one for the buyer’s prices, another for the seller’s prices.

### 3 The model

Let us denote by \(TN(\mu, \sigma^2)\) a normal random variable with mean \(\mu\) and variance \(\sigma^2\), truncated at \(x = 0\). Given a random sample from a truncated normal distribution with sample mean \(\bar{x}\) and empirical coefficient of variation \(c_n\), the likelihood equations are reduced to solving

\[
\frac{\alpha^2 + 1 - \alpha \lambda(\alpha)}{(\lambda(\alpha) - \alpha)^2} = 1 + c_n^2,
\] (2)

where \(\alpha = -\mu/\sigma\), and \(\lambda(\alpha) = n(\alpha)/(1 - N(\alpha))\) is the inverse of Mills ratio, in which \(n(x)\) and \(N(x)\) are the probability density function and the distribution function of the standard normal distribution. The maximum likelihood estimators are therefore given by

\[
\hat{\mu} = \frac{\pi \alpha}{\alpha - \lambda(\alpha)},
\] (3)
\[ \hat{\sigma} = \frac{\tau}{\lambda(\alpha) - \alpha}. \]  

(4)

Castillo (1994) shows that Equation (2) has a solution if \( c_n < 1 \) and provides a one-to-one map between \( c_n \) and \( \alpha \). The family of truncated normal distributions can thus be parameterized by the coefficient of variation and a scale parameter \( \sigma \). Hence, for any values of the scale parameter and the coefficient of variation \( (0 < cv < 1) \), we have a different model for non-negative data. Moreover, the exponential family of distributions appears as the limit case when \( cv = 1 \).

Let us denote by \( \text{Par}(\alpha, \sigma) \) the Pareto distribution having a probability density function:

\[ p(x; \alpha, \sigma) = \frac{\alpha \sigma^\alpha}{(x + \sigma)^{\alpha+1}}, \quad x > 0, \]  

(5)

where \( \sigma > 0 \) is a scale parameter and \( \alpha > 0 \) determines tail weight. It is important to note that, for the Pareto distribution, the coefficient of variation is a one-to-one function of \( \alpha \); it is always greater than 1, and can be any real number from 1 to infinity. The Pareto model can be parameterized by the coefficient of variation and the scale parameter according to

\[ \alpha = \frac{2cv^2 [x]}{c\alpha^2 [x] - 1}, \quad \sigma = (\alpha - 1) E [x]. \]  

(6)

The exponential distribution is a limit case when \( \alpha \) and \( \sigma \) tend to infinity with a linear restriction.

We now propose a simple model for non-increasing distributions with a mode at zero, useful for absolute values and for the positive and the negative sides of an assets-returns series. This model is the union of the truncated normal distribution, the exponential distribution and the Pareto distribution (for \( \alpha > 2 \)). The model parameterized by the coefficient of variation \( cv \) and a scale parameter \( \sigma \) is given by:

\[ TNP(cv, \sigma) = \begin{cases} 
TN(\mu, \sigma), & \text{if } cv < 1 \quad \mu, \sigma \text{ from (2), (3)}. \\
Exp(\sigma), & \text{if } cv = 1 \\
Par(\alpha, \sigma), & \text{if } cv > 1 \quad \alpha \text{ from (6)}. 
\end{cases} \]  

(7)

We then have a representative model for each level of the coefficient of variation between 0 and infinity. This provides a way to recognize the variation of tail behavior over time. The model is closed by changing the scale parameter \( \sigma \), that is, if \( x \sim TNP(cv, \sigma) \) then \( \lambda x \sim TNP(cv, \lambda \sigma) \), given \( \lambda > 0 \).

This model can be applied to a time series of asset returns by separately considering the positive and the negative subsamples or, if the symmetry hypothesis with respect to zero is accepted, by taking into account the absolute values of these returns. Moreover, a mean parameter can also be included if the symmetry seems restrictive, with respect to zero.
The exponential distribution is the limit case when the coefficient of variation goes to 1, at the same time, for the left and right sides. Because this situation is clearly simpler, it is of practical interest to check for \( cv = 1 \) with the test outlined in Castillo and Daoudi (2005). If we accept the null hypothesis of exponentiality, then we would say that the model has exponentially shaped tails; if we accept \( cv < 1 \), we would say that the tails have a normal shape; and if \( cv > 1 \), we would say that the model has Paretoian-shaped tails.

4 Main Results

We give a unified model, invariant for a scale parameter, which provides simple and representative distribution functions for each value of the coefficient of variation between zero and infinity. The model is built with a truncated normal distribution for \( cv < 1 \), an exponential distribution for \( cv = 1 \) and a Pareto distribution for \( cv > 1 \).

From the empirical point of view, for exchange rates between the euro and the US dollar, the model we propose provides satisfactory fits that improve on those obtained with the mixture of normal distributions. Daily and two-day returns show a coefficient of variation close to 1, which corresponds to the Laplace distribution in our model. The model also works well in cases of heavy-tailed distributions, as well as in cases where distributions are thin-tailed.

5 Conclusions

1. The coefficient of variation provides a new way for measuring tail heaviness for financial data. This measurement uses only two moments and is therefore more stable and more widely applicable than kurtosis. Heavy-tailed distributions correspond to coefficients of variation greater than 1 for threshold exceedances.

2. We introduce a simple model, invariant by scale parameter, in a one-to-one correspondence with all possible values of the coefficient of variation. This model provides a way of predicting and simulating marked variables. Our model unifies several useful finance models, such as the Laplace and Pareto distributions, together with certain new models.

3. From an economic point of view, the model we propose is reasonable because it is introduced from the spread between bidding-asking prices, using the idea of selective samples. It therefore provides an explanation of certain non-normality.
FIGURE 1. Fitted-probability density functions for GH ($\lambda$), $NIG$ ($\lambda = -0.5$) and normal distributions.

References


Marginal modelling of contingency tables with recursive logits and odds ratios

Manuela Cazzaro¹ and Roberto Colombi²

1 Dipartimento di Metodi Quantitativi per le Scienze Economiche ed Aziendali, Università di Milano-Bicocca, Italy, manuela.cazzaro@unimib.it
2 Dipartimento di Ingegneria Gestionale e dell’Informazione, Università di Bergamo, Italy, colombi@unibg.it

Abstract: In this work a new type of logits and odds ratios, which includes as special cases the continuation and the reverse continuation logits and odds ratios, is defined. We prove that these logits and odds ratios define a parameterization of the joint probabilities of a two way contingency table. Simple examples of testing equality and inequalities constraints on these logits and odds ratios are provided.

Keywords: Marginal models; recursive logits; generalized odds ratios.

1 Introduction

In this work a new type of logits and odds ratios (o.r.), which includes as special cases the continuation and the reverse continuation logits and odds ratios, is introduced. We prove that these logits and o.r., called recursive or nested, define a parameterization of the joint probabilities of a two way contingency table. As the logits are defined on marginal distributions, this work belongs to the marginal parameterizations context. This kind of parameterization has been developed during the last years, starting from the work of Glonek and McCullagh (1995) till more recent works as the one of Bartolucci, Colombi and Forcina (2005), (see also the papers therein quoted).

2 Recursive Logits and odds ratios

Let $A_1$, $A_2$ be two ordered categorical variables with categories in the sets $A_1 = \{a_{11}, a_{12}, \ldots, a_{1r}\}$ and $A_2 = \{a_{21}, a_{22}, \ldots, a_{2c}\}$ respectively. The numbering is assumed to be coherent with the order of the categories. By coherent complete hierarchy of sets $\mathcal{G}(A_1)$ a family of subsets of $A_1$ is intended. It is characterized by the following properties: (i) $A_1 \in \mathcal{G}(A_1)$; (ii) $\{a_{1i}\} \in \mathcal{G}(A_1)$, $i = 1, 2, \ldots, r$; (iii) if $P, Q \in \mathcal{G}(A_1)$ then $P \cap Q \in \{P, Q, \emptyset\}$; (iv) let $M_m$ be a non-minimal set according to the inclusion
order of \( G(A_1) \) then for \( m = 1, 2, \ldots, r - 1 \):

\[
M_m = B(a_{1i(m)}, 0) \cup B(a_{1i(m)}, 1), \quad B(a_{1i(m)}, 0), B(a_{1i(m)}, 1) \in G(A_1)
\]

\( i(m) = \sup\{i : a_i \in B(a_{1i(m)}, 0)\}, \quad i(m) + 1 = \inf\{i : a_i \in B(a_{1i(m)}, 1)\}. \)

Note that the last property, besides showing that the non-minimal sets of \( G(A_1) \) are \( r - 1 \), establishes a bijective mapping between the sets \( M_m \) and the first \( r - 1 \) categories of \( A_1 \). Furthermore from the same property follows that each set \( M_m \) contains contiguous categories.

The numbering of the non-minimal sets \( M_m \) is such that \( m > n \Rightarrow (M_n \supset M_m \vee \sup\{i : a_i \in M_n\} < \inf\{i : a_i \in M_m\}) \).

Analogously the family \( G(A_2) \) and its non-minimal sets \( N_n = B(a_{2j(n)}, 0) \cup B(a_{2j(n)}, 1), m = 1, 2, \ldots, c - 1 \) are defined.

Once \( G(A_1), G(A_2) \) have been specified, let us define, for each couple of non-minimal sets \( M_m, N_n, m = 1, 2, \ldots, r - 1, n = 1, 2, \ldots, c - 1 \), the following logits and log-odds ratios of recursive or nested type:

\[
\eta_{m1}^{A_1} = \ln \frac{P(A_1 \in B(a_{1i(m)}, 1))}{P(A_1 \in B(a_{1i(m)}, 0))}, \quad \eta_{m2}^{A_2} = \ln \frac{P(A_2 \in B(a_{2j(n)}, 1))}{P(A_2 \in B(a_{2j(n)}, 0))}
\]

(1)

\[
\eta_{m,n}^{A_1,A_2} = \ln \frac{P(A_1 \in B(a_{1i(m)}, 1), A_2 \in B(a_{2j(n)}, 1))}{P(A_1 \in B(a_{1i(m)}, 1), A_2 \in B(a_{2j(n)}, 0))}
\]

(2)

**Example.** *Continuation-recursive logits and log-o.r.* Let \( A_1 \) and \( A_2 \) have four levels each. Let the non-minimal sets of \( G(A_1) \) and \( G(A_2) \) be:

\[
M_1 = \{a_{11}, a_{12}, a_{13}, a_{14}\}, \quad M_2 = \{a_{12}, a_{13}, a_{14}\}, \quad M_3 = \{a_{13}, a_{14}\}
\]

\[
N_1 = \{a_{21}, a_{22}, a_{23}, a_{24}\}, \quad N_2 = \{a_{23}, a_{24}\}, \quad N_3 = \{a_{21}, a_{22}\}
\]

In this case the recursive logits of \( A_1 \) and \( A_2 \) are:

\[
\eta_{1}^{A_1} = \ln \frac{P(A_1 \in \{a_{12}, a_{13}, a_{14}\})}{P(A_1 \in \{a_{11}\})}, \quad \eta_{2}^{A_1} = \ln \frac{P(A_1 \in \{a_{13}, a_{14}\})}{P(A_1 \in \{a_{12}\})}
\]

\[
\eta_{1}^{A_2} = \ln \frac{P(A_2 \in \{a_{14}\})}{P(A_1 \in \{a_{13}\})}, \quad \eta_{2}^{A_2} = \ln \frac{P(A_2 \in \{a_{24}\})}{P(A_1 \in \{a_{22}\})}
\]

(3)

In particular note that \( A_1 \) has logits of continuation type. The log-odds ratios \( \eta_{m,n}^{A_1,A_2} \) for \( m = 1 \) and \( n = 1 \) is:

\[
\eta_{1,1}^{A_1,A_2} = \ln \frac{P(A_1 \in \{a_{12}, a_{13}, a_{14}\}, A_2 \in \{a_{23}, a_{24}\})}{P(A_1 \in \{a_{11}\}, A_2 \in \{a_{21}, a_{22}\})}
\]

(4)

Analogously the \( \eta_{m,n}^{A_1,A_2} \) are defined for \( m, n = 1, 2, 3, (m, n) \neq (1, 1) \).

In order to show that the marginal logits (1) and the recursive log-odds ratios (2) parameterize the joint probabilities \( \pi_{ij} \) of \( A_1 \) and \( A_2 \), we introduce the following probabilities for \( m = 1, 2, \ldots, r - 1, n = 1, 2, \ldots, c - 1 \):

\[
\mu_{m}^{A_1} = P(A_1 \in B(a_{1i(m)}, 1)) \quad \mu_{m}^{A_2} = P(A_2 \in B(a_{2j(n)}, 1))
\]

(3)

\[
\mu_{m,n}^{A_1,A_2} = P(A_1 \in B(a_{1i(m)}, 1), A_2 \in B(a_{2j(n)}, 1)).
\]

(4)
It can be shown that the previous probabilities define a \( m \)-parameterization, that is they are expected values of minimal sufficient statistics for the parameters of a saturated log-linear model. We will show that the mapping between the space of the recursive or nested logits and \( \log-o.r. \) (1)-(2) and the space of the probabilities (3)-(4) is a diffeomorphism. Hence the recursive or nested logits and \( \log-o.r. \) define a parameterization of the joint probabilities of \( A_1, A_2 \).

To each couple of sets \( (M_m, N_n) \), \( M_m \in \mathcal{G}(A_1) \) and \( N_n \in \mathcal{G}(A_2) \), \( m = 1, 2, ..., r - 1 \), \( n = 1, 2, ..., c - 1 \), the set of recursive or nested logits and \( \log-o.r. \) \( \Gamma(m, n) \) defined as:

\[
\Gamma(m, n) = \{ \eta_{m,n}^{A_1 A_2} \} \text{ if } m > 1, n > 1,
\]

\[
\Gamma(1, n) = \{ \eta_n^{A_2}, \eta_{1,n}^{A_1 A_2} \} \text{ if } n > 1,
\]

\[
\Gamma(m, 1) = \{ \eta_m^{A_1}, \eta_{m,1}^{A_1 A_2} \} \text{ if } m > 1,
\]

is associated. Note that the sets \( \Gamma(m, n) \) define a partition of the recursive or nested logits and \( \log-o.r. \) by connecting these logits and \( \log-o.r. \) only to one couple of sets \( (M_m, N_n) \). Then let \( \Gamma(m, n) = \bigcup_{(i,j) \in \text{pre}(m, n)} \Gamma(i, j) \), where \( \text{pre}(m, n) = \{(i, j) : (i < m) \lor (i = m, j < n)\} \), be the set of the recursive or nested logits and \( \log-o.r. \) associated to the couples of non-minimal sets \( (M_i, N_j) \) preceding \( (M_m, N_n) \) in the lexicographical order. Analogously we define the sets \( \Delta(m, n) \) and the sets \( \Delta(m, n) = \bigcup_{(i,j) \in \text{pre}(m, n)} \Delta(i, j) \) of the minimal sufficient statistics (3)-(4).

The main results are reported in the following Lemma and Theorem.

**Lemma 1** The mapping between the space of the minimal sufficient statistics, elements of \( \Delta(m, n) \cup \Delta(m, n) \) and the space of the recursive or nested logits and \( \log-o.r. \), elements of \( \Gamma(m, n) \cup \Gamma(m, n) \) is a diffeomorphism, for all \( m = 1, \ldots, r - 1 \) and \( n = 1, \ldots, c - 1 \).

We omit the proof (available from the authors) for lack of space.

**Theorem 1** The recursive or nested logits and \( \log-o.r. \) parameterize the joint distribution of \( A_1 \) and \( A_2 \).

**Proof.** The mapping between the space of the parameters of the saturated log-linear model of the joint distribution of \( A_1 \) and \( A_2 \) and the space \( \Delta = \Delta(r - 1, c - 1) \cup \Delta(r - 1, c - 1) \) of the minimal sufficient statistics is a diffeomorphism as the multinomial probability function described by the \( \pi_{ij} \) belongs to the exponential family. Lemma 1 establishes the diffeomorphism between the space of the minimal sufficient statistics, elements of \( \Delta \), and the space of the recursive or nested logits and \( \log-o.r. \), elements of \( \Gamma(1, n) \cup \Gamma(1, n) \). Hence the mapping between the space of the log-linear model parameters and the space of the recursive or nested logits and \( \log-o.r. \), elements of \( \Gamma \), is a diffeomorphism.

### 3 Hypothesis testing example

Let \( \eta \) be the vector of the recursive logits and \( \log-o.r. \) (1)-(2). In this context it is interesting to test equality and inequality constraints on the parameter...
vector $\eta$ under multinomial sampling. Constraints defined by equalities may be useful for testing independence in sub-tables of the contingency table or marginal homogeneity. Inequality constraints can verify if the two variables are positively dependent. The way of testing linear equality and inequality constraints on $\eta$ is analogous to that described by Colombi and Forcina (2001). Here we remind only that in presence of inequality constraints the likelihood ratio test statistics has an asymptotic chi-bar square distribution.

To give an example we consider the data of Table 1, where customers, users of a machine tool, are asked to evaluate their satisfaction (unsatisfied (U), satisfied (S), really satisfied (RS), extremely satisfied (ES)) regarding the customer service and the training given by the company that sells the machine tool. Note that the non-minimal sets of the variables are: $M_1 = N_1 = \{U, S, RS, ES\}$, $M_2 = N_2 = \{RS, ES\}$, $M_3 = N_3 = \{U, S\}$.

<table>
<thead>
<tr>
<th>Customer service \ Training</th>
<th>U</th>
<th>S</th>
<th>RS</th>
<th>ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>9</td>
<td>3</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>S</td>
<td>10</td>
<td>4</td>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>RS</td>
<td>22</td>
<td>10</td>
<td>81</td>
<td>31</td>
</tr>
<tr>
<td>ES</td>
<td>25</td>
<td>4</td>
<td>37</td>
<td>97</td>
</tr>
</tbody>
</table>

Testing different models against the unrestricted alternative hypothesis gives the following results: (i) positive recursive log-o.r. model: $G^2 = 4.83$, $p$-value $= 0.3672$; (ii) positive recursive log-o.r. and $\eta_1^A = \eta_2^A$, $\eta_1^B = \eta_2^B$ model: $G^2 = 3.97$, $p$-value $= 0.1372$; (iii) positive recursive log-o.r. and homogeneity of the marginals model: $G^2 = 14.43$, $p$-value $= 0.0024$.

References


Association structure and ranking of emergence times of permanent teeth: modelling the covariance matrix with interval-censored data

Silvia Cecere, Alejandro Jara and Emmanuel Lesaffre

1 Biostatistical Centre, Katholieke Universiteit Leuven, Kapucijnenvoer 35, 3000 Leuven, Belgium

Abstract: We examine a Bayesian approach for the modelling of the covariance matrix of a multivariate normal distribution where the covariance structure is allowed to depend on covariates and in the presence of interval-censoring. The motivating example is taken from dental research where the association between emergence times of different teeth in the mouth is explored. Analyzing the association of emergence times of permanent teeth is challenging because the emergence times most often are left- or right- or interval-censored. There is interest in examining the impact of covariates on the association among emergence times and in establishing the preferred orderings of emergence and the dependence of these preferred orderings on covariates. While the presence of interval censoring pleads for an MCMC approach employing the data augmentation algorithm, the fact that variance-covariance matrices necessarily need to be positive definite complicates the computations considerably. In this paper, we used the modified Cholesky decomposition (Pourahmadi, 1999) which ensures the pd condition. The procedure is illustrated on the analysis of emergence times of permanent teeth from one quadrant in the mouth on children from the longitudinal Signal Tandmobiel® project.

Keywords: Bivariate survival; covariance matrix; positive definiteness; data augmentation; interval-censored.

1 Introduction

From a dental point of view, it was of interest to examine the impact of caries experience and gender on the association structure of emergence times of permanent teeth and to determine the most preferred sequences of teeth emergence. To this end, we employed the Signal Tandmobiel® data, a 6 year longitudinal oral health study started in Flanders (Belgium) in 1996 in which 4468 children were annually examined on pre-scheduled visits by sixteen trained dental examiners in a mobile dental clinic. At each visit, the tooth emergence status was recorded. Caries experience was measured by
112 Modelling of the covariance matrix

the dmft-index which is the sum of the deciduous teeth which are either decayed, or missing due to extraction because of caries, or filled. Since children were annually examined, the emergence time was recorded in an interval-censored manner. That is, either the tooth had emerged before the first examination (left-censored), emerged in-between two examinations (interval-censored), or didn’t emerge by the last examination (right-censored). In order to explore the impact of covariates on the association structure, we needed to model the covariance matrix of the unknown, latent, emergence times. Modeling a covariance matrix as a function of covariates is complex due to the requirement that it should be positive-definite (pd). A variety of approaches has been suggested to deal with the pd-condition. Further, in statistical modelling an important aspect is the interpretability of the model parameters. The approach suggested by Pourahmadi (1999) offers a trade-off between dealing with the pd-condition and interpretability of the regression coefficients of the covariates. More specifically, he proposed to use the modified Cholesky decomposition of the covariance matrix. In effect, given a symmetric matrix $\Sigma$, $\Sigma$ is positive definite if and only if there exists a unique unit lower triangular matrix $T$ and a unique diagonal matrix $D$ having positive diagonal entries, such that

$$\Sigma = TDT^t.$$ 

Thus, given $T$ and $D$ as required, the resulting matrix $\Sigma$ is pd. In addition, the elements of $T$ can be interpreted as regression coefficients of conditional regression models, while the diagonal elements of $D$ represent conditional variances. Hence, this decomposition allows to introduce covariates very easily through the unconstraint entries of the $T$ matrix and the logarithm of the diagonal entries of the $D$ matrix without any concern for the pd condition. Modelling jointly the marginal and the association structure allowed us to establish the most preferred emergence rankings and to see how much the ranking depends on these covariates. Establishing the prevalence of each emergence ranking is not trivial due to the large number of possible rankings. We explored the seven permanent teeth of the right maxilla quadrant, which are denoted as $1x$, where $x = 1, \ldots, 7$ in European notation. The incisors are denoted as 11 and 12, the canine as 13, the pre-molars as 14 and 15, the molars as 16 and 17. Hence, there were $7! = 5040$ possible rankings to consider.

1.1 Methodology

Assume that the response vector (of true emergences) for the $i$-th subject follows a multivariate normal distribution, i.e.

$$Y_i \sim N_p(x_i, \beta_i, \Sigma_i), \ (i = 1, \ldots, n)$$

where $Y_i$ is a $p \times 1$ vector of responses, $\beta = (\beta_1^T, \ldots, \beta_q^T)^T$, with $\beta_s$ a $q \times 1$ vector of regression coefficients corresponding to the $s$th response, and $x_i$
is a $p \times (q \times p)$ design matrix given by

$$x_i = I_{pp} \otimes x_i^T,$$

where $I_{pp}$ is the $p \times p$ identity matrix, $\otimes$ the Kronecker product and $x_i^T = (1, x_{i1}, \ldots, x_{iq-1,i})$. Further, $\Sigma_i$ is the covariance matrix of the $i$th subject. A natural parametrization of the covariance matrix, discussed by Barnard et al. (2000), is to use the variance-correlation decomposition. More specifically, this decomposition implies that $\Sigma_i = \Gamma_i R_i \Gamma_i$, where $\Gamma_i$ is the diagonal matrix of the standard deviations and $R_i$ the correlation matrix. Covariates can be introduced through the entries of the $\Gamma$ matrix and the entries of the correlation matrix. For the diagonal elements of $\Gamma_i$, it is common to use the log link, i.e. $\log(\sigma_{si}^2) = x_i^T \lambda_s$. With respect to the correlation coefficients, it is useful to model its Fisher’s (or an analogous) transforms. That is, if $\rho_{st,i}$ is the $(s, t)$th entry of $R_i$ then $\rho_{st,i} = \tanh(x_i^T \psi(s, t))$. Fisher’s transform removes the implied range constraints ($\rho \in [-1, 1]$) on the covariates. However, it is not sufficient to ensure positive definiteness of $R_i$ when $p > 2$. In Figure 1 we display the minimum eigenvalue of a $3 \times 3$ correlation matrix where the Fisher’s transform of the correlations depends linearly on a continuous covariate $x$. Clearly, positive definiteness is violated in some regions of the support of the covariate. But, even if $R_i$ were positive definite in the range of the observed $x$-values of the sample, there is no guarantee that the correlation matrix remains positive definite for future values of $x$. As mentioned before, Pourahmadi (1999) used the modified Cholesky decomposition of $\Sigma_i^{-1}$ to propose a statistically interpretable unconstrained parametrization of the covariance matrix. In effect, suppose $\beta$ is given and $\Sigma_i \equiv \Sigma$. For simplicity reasons, we temporarily omit the dependence on the subject. Let $\hat{Y}_s$ be the linear least-squares predictor of $Y_s$, the $s$th component of $Y$, on its predecessors $Y_{s-1}, \ldots, Y_1$ and $\epsilon_s = \hat{Y}_s - Y_s$ be its prediction error with variance $\tau_s^2 = \text{var}(\epsilon_s)$. Thus,

$$Y_s = \hat{Y}_s + \epsilon_s = x_i^T \beta_s + \sum_{j=1}^{s-1} \phi_{sj}(Y_j - x_j^T \beta_j) + \epsilon_s,$$

which is obtained by standard regression arguments. The components of this regression models are related to the elements of $T$ and $D$. More specifically, $D = \text{diag}(\tau_1^2, \ldots, \tau_p^2)$ and $-\phi_{st}$ is the $(s, t)$ entry of the matrix $T$. Hence, following Pourahmadi (1999) it is proposed here to model the dependence of the association structure on covariates as follows:

$$\log(\tau_{si}^2) = x_i^T \lambda_s,$$

$$\phi_{st,i} = x_i^T \gamma_{st},$$

where $\lambda_s$ ($s = 1, \ldots, p$) and $\gamma_{st}$ ($s = 2, \ldots, p; t = 1, \ldots, (p - 1)$) are vectors of regression coefficients for the log-transformed conditional variances and
the dependence parameters respectively. Expressions (2) and (3) can be generalized by allowing different covariates. Summarized, the parameters to be estimated in the model given the true, latent, emergence times are:

- The regression parameters of the marginal distribution ($\beta$).
- The regression parameters of the dependence parameters ($\gamma$).
- The regression parameters of the conditional variances ($\lambda$).

For the interpretability of the parameters of the Modified Cholesky decomposition, a natural ordering of the components of the response vector is assumed. Such a natural ordering is present in longitudinal studies or time series. However, in our dental example, we cannot claim that there is a natural ordering of the emergence times. In fact, one of our goals is to look for the preferred rankings. Thus, we considered a working sequence, the one given by the mean emergence times previously established by our group (see e.g., Leroy et al. (2003)). Based on this, we have ordered the components of the teeth emergence vector $Y$ as follows: 16-11-12-14-13-15-17.

2 Modelling multivariate normal interval censored data

For the estimation of the regression parameters, we have followed the approach of Daniels and Pourahmadi (2002). Hence, we refer to that paper...
for more technical details. Briefly, the prior distributions for the parameter vectors $\beta$, $\lambda$ and $\gamma$ associated to the time emergence times are

$$
\beta \sim N(\beta^*, \Sigma_\beta), \quad \lambda \sim N(\lambda^*, \Sigma_\lambda), \quad \gamma \sim N(\gamma^*, \Sigma_\gamma).
$$

The normal priors for $\beta$ and $\gamma$ are conditionally conjugate to the full conditionals of the respective parameter vectors. Indeed, the full conditionals for $\beta$ and $\gamma$ are normals, which allows straightforward Gibbs sampling. The parameter vector $\lambda$ can be sampled using a Random Walk Metropolis with a normal proposal. An extra Gibbs step using the Data Augmentation algorithm allows to generate easily the latent emergence times, given the observed interval emergence times. Basically, at each iteration of the MCMC chain, all latent observations are imputed from their full conditional distribution and then the parameter vector is updated based on the complete imputed sample. Since the observations are independent given $\beta$, $\gamma$ and $\lambda$ and assuming that censoring occurs non-informatively (as in the Signal Tandmobiel study), the full conditional distribution of $Y_i$ given all the remaining parameters and the data are multivariate normal truncated to the hyper-cube determined by the intervals where the emergence took place. In order to sample from this full conditional distribution, at each iteration of the Gibbs sampler, we sampled iteratively the components of $Y_i$.

The prevalence of all possible sequences of emergence was calculated for each gender $\times$ dmft combination separately. Let a sequence of emergence times be denoted as $i_1i_2i_3i_4i_5i_6i_7$, then the prevalence of this sequence was determined by

$$
P(Y_{i_1} < Y_{i_2} < Y_{i_3} < Y_{i_4} < Y_{i_5} < Y_{i_6} < Y_{i_7}) = \int_S \phi(x|\mu, \Sigma) dx,
$$

where $\phi(x)$ denotes the multivariate normal density with $\mu$ and $\Sigma$ set equal to the estimated corresponding mean and covariance matrix for the given gender $\times$ dmft combination, and $S = \{(y_{i_1}, y_{i_2}, y_{i_3}, y_{i_4}, y_{i_5}, y_{i_6}, y_{i_7}) \in R^7 | y_{i_1} < y_{i_2} < y_{i_3} < y_{i_4} < y_{i_5} < y_{i_6} < y_{i_7}\}$. This highly-dimensional integral over the unbounded region $S$ was numerically calculated using Quasi-Monte Carlo Integration techniques.

## 3 Results

### 3.1 Multivariate Normal Model

The posterior means and the 95% HPD intervals of the regression parameters of the mean structure $\beta$ are shown in Table 1. It can be seen that the permanent teeth emerge earlier in girls than in boys, corroborating the findings of others. The present analysis shows that the emergence of all teeth is accelerated when there is (more) caries on the deciduous teeth, because of a negative regression coefficient for dmft. But only for teeth 13 (canine), 14 and 15 (pre-molars) the 95% HPD-interval does not encompass 0. These results confirm what has been reported by Leroy et al. (2003 a).
TABLE 1. Posterior means (95 % HPD) of regression coefficients ($\beta$) for the marginal part of the multivariate normal model.

<table>
<thead>
<tr>
<th>Tooth</th>
<th>Intercept</th>
<th>Gender (1=Girls)</th>
<th>$dmft \times 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_1$</td>
<td>16</td>
<td>6.44</td>
<td>-0.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(6.40,6.49)</td>
<td>(-0.19,-0.07)</td>
</tr>
<tr>
<td>$Y_2$</td>
<td>11</td>
<td>7.11</td>
<td>-0.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(7.07,7.16)</td>
<td>(-0.30,-0.19)</td>
</tr>
<tr>
<td>$Y_3$</td>
<td>12</td>
<td>8.30</td>
<td>-0.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(8.25,8.35)</td>
<td>(-0.45,-0.33)</td>
</tr>
<tr>
<td>$Y_4$</td>
<td>14</td>
<td>11.03</td>
<td>-0.41</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(10.96,11.10)</td>
<td>(-0.49,-0.33)</td>
</tr>
<tr>
<td>$Y_5$</td>
<td>13</td>
<td>11.64</td>
<td>-0.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(11.55,11.71)</td>
<td>(-0.73,-0.54)</td>
</tr>
<tr>
<td>$Y_6$</td>
<td>15</td>
<td>11.89</td>
<td>-0.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(11.80,11.99)</td>
<td>(-0.49,-0.27)</td>
</tr>
<tr>
<td>$Y_7$</td>
<td>17</td>
<td>12.38</td>
<td>-0.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(12.28,12.49)</td>
<td>(-0.45,-0.19)</td>
</tr>
</tbody>
</table>

Although the parameters in the Modified Cholesky decomposition are statistically interpretable from a conditional point of view, it is not immediately clear what the impact of these covariates is on the marginal variances or on the correlation coefficients. Both the marginal variances and the correlation coefficients are highly non-linear functions of the covariates. Since we are interested in the marginal variance-covariance matrix a simple way to obtain a posterior estimate is to sample the posterior predictive distribution given a priori chosen gender $\times dmft$-combinations. As an illustration, Table 2 shows the posterior mean of some the correlation coefficients as a function of gender and $dmft$.  

TABLE 2. Posterior mean of correlation coefficients by gender $\times dmft$

<table>
<thead>
<tr>
<th>Node</th>
<th>$dmft$</th>
<th>Girls</th>
<th>Boys</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>$\rho_{45}$</td>
<td>0.74</td>
<td>0.72</td>
<td>0.61</td>
</tr>
<tr>
<td>$\rho_{46}$</td>
<td>0.79</td>
<td>0.77</td>
<td>0.69</td>
</tr>
<tr>
<td>$\rho_{47}$</td>
<td>0.65</td>
<td>0.63</td>
<td>0.53</td>
</tr>
<tr>
<td>$\rho_{56}$</td>
<td>0.65</td>
<td>0.62</td>
<td>0.50</td>
</tr>
<tr>
<td>$\rho_{57}$</td>
<td>0.59</td>
<td>0.58</td>
<td>0.51</td>
</tr>
</tbody>
</table>
We see that for all teeth the correlation coefficient decreases with increasing $dmft$-index.

### 3.2 Prevalence of Emergence

Our analysis showed that the ranking according to the median sequence is also the most prevalent ranking for almost all gender × $dmft$-combinations with prevalence varying between 10% and 18%. An important finding of our model is that caries distorts the emergence process. The proof of this distortion is seen in Table 3. Firstly, we reported the number of rankings needed to cover 90% of the probabilities of the 7! possible rankings. For this, we first ranked the rankings in decreasing prevalence. It is seen in Table 3, that with increasing $dmft$-index the necessary number of rankings increases. A second measure is given by Shannon’s entropy coefficient. Clearly, the entropy coefficient increases with $dmft$-index indicating the ranking probabilities tend to show a more uniform profile when there is more caries in the deciduous teeth. For more information on the actual probabilities of the most prevalent rankings in the Flemish population we refer to Cecere et al. (2006).

<table>
<thead>
<tr>
<th>$dmft$</th>
<th>N</th>
<th>Entropy</th>
<th>N</th>
<th>Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>27</td>
<td>3.226</td>
<td>24</td>
<td>3.205</td>
</tr>
<tr>
<td>1</td>
<td>27</td>
<td>3.229</td>
<td>25</td>
<td>3.189</td>
</tr>
<tr>
<td>5</td>
<td>32</td>
<td>3.303</td>
<td>30</td>
<td>3.315</td>
</tr>
<tr>
<td>10</td>
<td>51</td>
<td>3.720</td>
<td>47</td>
<td>3.700</td>
</tr>
<tr>
<td>15</td>
<td>96</td>
<td>4.352</td>
<td>83</td>
<td>4.281</td>
</tr>
</tbody>
</table>

### 3.3 Concluding Remarks

In this paper we have illustrated that the approach of Pourahmadi (1999) is useful for modelling covariance matrices. Here we have exemplified this approach for a relatively high dimension and in a completely different application area. Further, the interpretability of the parameters is satisfactory from a conditional view point. We were mainly interested in the effect of covariates on the marginal association structure. At this point it may be argued that there are other decompositions of covariance matrices which ensure the pd-condition but at the cost of lack of statistical interpretation. We have calculated the prevalence of the most preferred rankings of
emergence. Our analysis is one of the very few that examined the ranking of emergence times on such a large and detailed dental data set and with well justified statistical technique avoiding as much as possible ad hoc procedures (like the mid-point approach).

Acknowledgments: Special Thanks to Roos Leroy for the stimulating discussions prior to preparing the paper and the careful reading of the manuscript. Further, the authors have been partly supported by the Research Grants OT/00/35 and OT/05/60, Catholic University Leuven. The second author has also been partly supported by the KUL-PUC bilateral (Belgium-Chile) grant. The authors also acknowledge the partial support from the Interuniversity Attraction Poles Program P5/24 – Belgian State – Federal Office for Scientific, Technical and Cultural Affairs. Data collection was supported by Unilever, Belgium. The Signal Tandmobiel® study comprises following partners: D. Declerck (Dental School, Catholic University Leuven), L. Martens (Dental School, University Ghent), J. Vanobbergen (Oral Health Promotion and Prevention, Flemish Dental Association), P. Bottenberg (Dental School, University Brussels), E. Lesaffre (Biostatistical Centre, Catholic University Leuven) and K. Hoppenbrouwers (Youth Health Department, Catholic University Leuven; Flemish Association for Youth Health Care).

References


A log-linear randomized-response model to account for cheating

M.J.L.F. Cruyff, A. van den Hout, P.G.M. van der Heijden and U. Böckenholt

1 Utrecht University, The Netherlands, m.crujff@fss.uu.nl
2 MRC, Cambridge, United Kingdom, ardo.vandenhout@mrc-bsu.cam.ac.uk
3 McGill University, Montreal, Canada, ulf.bockenholt@mcgill.ca

Abstract: Randomized response (RR) is an interview technique designed to eliminate response bias when sensitive questions are asked. In RR the answer depends to a certain degree on the outcome of a randomizing device. Although RR elicits more honest answers than direct questions, respondents do not always follow the instructions and in the sense that they answer no regardless of the outcome of the randomizing device. In this paper we present a log-linear randomized-response model that accounts for this kind of cheating. The main results of this model are (1) an estimate of the probability of cheating; (2) log-linear parameters estimates describing the associations between RR variables and; (3) prevalence estimates of the sensitive behavior that are corrected for cheating. We illustrate the model with an example.

Keywords: Randomized response; log-linear model; cheating parameter.

1 Introduction

Most people are reluctant to publicly answer questions about sensitive topics, like drug or alcohol (ab)use, sexuality or anti-social behavior. As a result, respondents may refuse to give the embarrassing answer and the stigmatizing behavior is often underreported. Randomized Response (RR) is an interview technique that is especially developed to eliminate this kind of evasive response bias (Warner, 1965). In RR the answer to the sensitive question is to a certain extent determined by a randomizing device, like a pair of dice or the draw of card. Since the outcome of the device is known only to the respondent, confidentiality is guaranteed. A meta-analysis shows that RR yields more valid prevalence estimates than direct-questioning designs (Lensvelt-Mulders, Hox, Van der Heijden and Maas, 2005).

Despite the protection of the respondents’ privacy, RR does not completely eliminate the response bias. Several studies have shown that the RR design is susceptible to cheating, in the sense that the respondent does not answer in accordance with the outcome of the randomizing device. For example, in an experimental setting, Edgell, Himmelfarb and Duchan (1982) used an
RR design of which the outcomes of the randomization procedure were fixed in advance. It appeared that to a question about having experiences with homosexuality, 25% of the respondents who had to answer *yes* by design cheated. In another study, Van der Heijden, Van Gils, Bouts and Hox (2000) applied different interview techniques to subjects who were identified as having committed social welfare fraud. Although the RR condition elicited more admittances of fraud than direct-questioning or computer-assisted self-interviews, a substantial proportion of the subjects still denied having committed fraud. Finally, after completing a computer-assisted RR survey, most respondents stated they had found it hard to give a false *yes* response and some admitted having cheated (Boeije and Lensvelt-Mulders, 2002). Böckenholt and van der Heijden (2004) propose an item randomized-response model to correct for cheating. In this paper we present a log-linear modeling approach to account for cheating. Chen (1989) and Van den Hout and Van der Heijden (2004) have presented log-linear randomized-response (LLRR) models to study the associations between RR variables. In the present paper a log-linear randomized cheating (LLRRC) model is specified by the introduction of an extra parameter into the LLRR model to account for cheating. The three main results of LLRRC model are: (1) an estimate of the probability of cheating; (2) log-linear parameter estimates describing the associations between RR variables and; (3) prevalence estimates of the sensitive behavior that are corrected for cheating. The model is illustrated with an example from the Social Welfare Survey conducted in 2000 in the Netherlands (Van Gils, Van der Heijden, Bouts and Hox, 2000, and Lensvelt-Mulders, Van der Heijden, Laudy and Gils, 2006).

The outline of this paper is as follows. In section 2 we present the questions and the RR design used in the Social Welfare Survey. Section 3 introduces the general RR model and shows that identification problems arise when a cheating parameter is introduced. Then the log-linear randomized-response (LLRR) model is introduced as a reparametrization of the general RR model, followed by the log-linear randomized-response cheating (LLRRC) model. This model is not overparametrized because the cheating parameter is incorporated into the model at the expense of the highest-order interaction parameter. We then present the results of the example from the Social Welfare Survey. The example is followed by an investigation of the robustness of the parameter estimates against violations of model assumptions.

## 2 Survey Data

In the Netherlands employees are insured under various social welfare acts against the loss of income due to redundancy, invalidity or sickness. Recipients of financial benefits have to comply with the rules and regulations that are stipulated in these acts. Violation of the rules is considered fraud and can have serious repercussions for the offender. In 2000 the Dutch De-
partment of Social Affairs conducted a nationwide survey to monitor the degree of noncompliance with respect to these rules. From the 2000 Social Welfare Survey we present an examples of the LLRRC model. The example concerns the following four questions that were asked to 1,308 beneficiaries of the Disability Benefit Act (DBA):

1. Has a doctor or specialist ever told you that the symptoms your disability classification is based upon have decreased without you informing the Department of Social Services of this change?

2. At a Social Services check-up, have you ever acted as if you were sicker or less able to work than you actually are?

3. Have you yourself ever noticed an improvement in the symptoms causing your disability, for example in your present job, in volunteer work or the chores you do at home, without informing the Department of Social Services of this change?

4. For periods of any length at all, do you ever feel stronger and healthier and able to work more hours without informing the Department of Social Services of this change?

The answers to these questions are denoted by the variables $D$, $E$, $F$ and $G$. The questions were all answered according to the Kuk design (Kuk, 1990 and, Van der Heijden, Van Gils, Bouts and Hox, 2000). In this particular RR design the respondent is given two decks with red and black playing cards. One deck contains 80% red cards and 20% black cards and is called the yes deck. The other deck contains 80% black cards and 20% red cards and is called the no deck. Each time a sensitive question is asked, the respondent draws one card from both decks and answers the question by naming the color of the card from the deck corresponding to the true answer. So, if the true answer is yes, the respondent names the color of the card from the yes deck, and if the true answer is no, the respondent names the color of the card from the no deck. The vector with the observed-response frequencies $yyyy, yyyn, \ldots, nnnn$ is given by $n^* = (43, 22, 10, 34, 20, 31, 40, 93, 30, 29, 40, 91, 60, 86, 146, 533)^t$.

3 The Log-Linear Randomized-Response Cheating Model

3.1 A general Randomized-Response Model

Consider a multivariate RR design with $K$ dichotomous sensitive questions 1, 2, \ldots, $K$. Let the true responses be denoted by the random variables $A, B, \ldots$, with realizations $a, b, \ldots \in \{1 \equiv yes, 2 \equiv no\}$, and let the random variable $X$ denote the true-responses profiles $A = a, B = b, \ldots$. Analogously, let $X^*$ denote the observed-response profiles. A general RR model is given by

$$\pi^* = P_K \pi,$$

(1)
where \( \pi = (\pi_1, \ldots, \pi_D)^t \) denotes the true-response profile probabilities. The multivariate matrix \( P_K \), with elements \( p_{ij} = P(X^* = i|X = j), \) \( i, j \in \{1, \ldots, D\} \) denoting the conditional misclassification probabilities, is a transition matrix given by the Kronecker products \( P_1 \otimes P_1 \otimes \cdots \otimes P_1 \) of the univariate transition matrix \( P_1 \). For the Kuk design

\[
P_1 = \begin{pmatrix}
    p_{11} & p_{12} \\
    p_{21} & p_{22}
\end{pmatrix} = \begin{pmatrix}
    8/10 & 2/10 \\
    2/10 & 8/10
\end{pmatrix}.
\]

(2)

### 3.2 Boundary solutions

The general RR model sometimes exhibits a lack of fit. We refer to such a result as a boundary solution, because in the general RR model a lack of fit is characterized by one or more parameter estimates on the boundary of the parameter space (Van den Hout and Van der Heijden, 2002). A lack of fit is a somewhat unexpected result, because the general RR model is a saturated model in the sense that the number of independent parameters equals the number of independent observed relative frequencies. There are two potential causes for boundary solutions to occur.

Define the probability of observing a \textit{yes} response given that the prevalence of the sensitive characteristic is zero as the chance level. Boundary solutions occur when the relative frequency of a response or response profile is below chance level. One potential cause for a boundary solution is RR sampling variation, by which we mean the random fluctuation in the sample average of the number of red cards. If the prevalence of the sensitive characteristic is zero and less red cards are drawn from the \textit{no} deck than the expected 20\%, the relative frequency of \textit{yes} responses will be below chance level. The other potential cause for a boundary solution is cheating. If some of the respondents who have drawn a red card answer \textit{no} instead of the required \textit{yes}, the relative frequency of \textit{yes} responses will drop below chance level if the true prevalence is (near) zero.

### 3.3 A general Randomized-Response Cheating Model

Cheating can be modeled in the general RR model by the introduction of a cheating parameter \( \theta \),

\[
\pi^* = (1 - \theta)P_K \pi + \theta \nu = Q_K \pi
\]

(3)

where \( \theta \) denotes the probability of cheating, \( \nu \) is the \( D \)-dimensional vector \((0, \ldots, 1)^t\), and the transition matrix \( Q_K \) has elements

\[
q_{ij} = \begin{cases}
    (1 - \theta)p_{ij} & \text{for } i \neq D, j \in \{1, \ldots, D\} \\
    (1 - \theta)p_{ij} + \theta & \text{for } i = D, j \in \{1, \ldots, D\}
\end{cases}
\]

(4)
Notice however that model (3) is not identified, in the model there is one more parameter than the observed number of independent relative response frequencies.

3.4 The Log-Linear Randomized-Response Cheating Model

The log-linear randomized-response (LLRR) model is presented by Chen (1989) in the context of misclassification of categorical data and is further developed by Van den Hout and Van der Heijden (2004). Consider the true-response variables $A, B$ and $C$. The saturated log-linear model $[ABC]$ is given by

$$\log \pi_{abc} = \lambda_0 + \lambda_a^A + \lambda_b^B + \lambda_c^C + \lambda_{ab}^{AB} + \lambda_{ac}^{AC} + \lambda_{bc}^{BC} + \lambda_{abc}^{ABC},$$  

where the $\lambda$ terms are constrained to sum to zero over any subscript. The variables $A, B$ and $C$ are not directly observed, but are linked with the observed data through the misclassification probabilities. The log-linear model linking the true-responses with the observed data is obtained by defining $\eta_j = \log \pi_j$, $j \in \{1, \ldots, D\}$, and $\eta = (\eta_1, \ldots, \eta_D)^t = M\lambda$, where $M$ is the $D \times r$ design matrix of the model and $\lambda$ is the $r \times 1$ parameter vector, for $r \in \{1, 2, \ldots, D\}$.

Model (5) is easily adapted to accommodate cheating by replacing the elements $p_{ij}$ of transition matrix $P_K$ by the elements $q_{ij}$ of transition matrix $Q_K$ defined in (4). However, as for the general RR model the incorporation of a cheating parameter leads to an overparameterized model. We solve this problem by constraining the highest-order interaction parameter to zero. In a design with $K$ variables, constraining the $K$-factor interaction parameter ensures that the hierarchical structure of the model is preserved. In the case of the three variables $A, B$ and $C$, the unconstrained LLRRRC model (i.e. the model with all parameters except the $K$-factor interaction parameter), $\theta, [AB, AC, BC]$ is given by

$$\log \pi_{abc} = \lambda_0 + \lambda_a^A + \lambda_b^B + \lambda_c^C + \lambda_{ab}^{AB} + \lambda_{ac}^{AC} + \lambda_{bc}^{BC}.$$  

The kernel of the log-likelihood of the LLRRRC model is given by

$$\ell(\lambda, \theta | n^*, P_K) = \sum_{i=1}^D n_i^* \log \left( \sum_{j=1}^D q_{ij} e^{\eta_j} \right),$$  

where the terms $q_{ij}$ refer to the elements of the transition matrix $Q_K$ defined in (4). Apart from the transition matrix, model (6) is identical to the LLRR model, with $\eta_j = \log \pi_j$, $j \in \{1, 2, \ldots, D\}$, $M$ the $D \times r$ design matrix, $\lambda$ the $r \times 1$ parameter vector, and $\eta = M\lambda$. In the unconstrained LLRR model $r \in \{1, 2, \ldots, D - 1\}$, and the dimensions of the parameter vector $\lambda$ and the design matrix $M$ can be further reduced by constraining interaction parameters to be zero or equal to each other.
4 Example

Table 1 presents a summary of the process of model selection. The table reports the likelihood-ratio statistics $L^2$ that were obtained from fitting various LLRR models and LLRRC models. The table also presents the cheating parameter estimates for the LLRRC models.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\hat{\theta}$</th>
<th>$L^2$</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>H0: $[DEFG]$</td>
<td>-</td>
<td>37.3</td>
<td>0</td>
</tr>
<tr>
<td>H1: $[DEF, DEG, DFG, EFG]$</td>
<td>-</td>
<td>38.9</td>
<td>1</td>
</tr>
<tr>
<td>H2: $\hat{\theta}, [DEF, DEG, DFG, EFG]$</td>
<td>.15 (.03)</td>
<td>7.1</td>
<td>0</td>
</tr>
<tr>
<td>H3: $\hat{\theta}, [DE, DF, DG, EF, EG, FG]$</td>
<td>.13 (.05)</td>
<td>7.1</td>
<td>4</td>
</tr>
<tr>
<td>H4: $\hat{\theta}, [DE, DF, DG, EF, EG, FG]$ $^a$</td>
<td>.13 (.05)</td>
<td>36.2</td>
<td>9</td>
</tr>
<tr>
<td>H5: $\hat{\theta}, [DE, EF, FG]$ $^b$</td>
<td>.15 (.03)</td>
<td>8.4</td>
<td>8</td>
</tr>
<tr>
<td>H6: $\hat{\theta}, [D, E, F, G]$</td>
<td>.27 (.02)</td>
<td>82.9</td>
<td>10</td>
</tr>
</tbody>
</table>

$^a$ equality constraints on all interaction parameters
$^b$ equality constraints $\lambda_{DE} = \lambda_{EF}$

The models H0 and H1 are LLRR models. The saturated LLRR model H0 fit poorly. In the model H1 the highest-order interaction parameter $\lambda_{DEFG}$ is constrained to zero. The slight deterioration in fit indicates the absence of substantial $K$-factor interaction in the data when cheating is not taken into account. Models H2 to H6 are LLRRC models. The unconstrained LLRRC model H2 has likelihood-ratio statistic of 7.1. Elimination of all 3-factor interaction parameters in model H3 does not affect the fit. The model H5, with equality of the interaction parameters $\lambda_{DE}$ and $\lambda_{EF}$, is the most parsimonious model, with an estimated probability of cheating of 0.15. Model H4 and model H6 illustrate that the fit deteriorates when more constraints are imposed.

The associations between the individual variables are presented in Table 2 for the LLRRC model H5. The odds ratios suggest that associations between the variables $D$ and $E$ and the variables $E$ and $F$ in the Health example are very strong, implicating that fraud with respect variable to $E$ is almost always associated with fraud with respect to variables $D$ and $F$. The variables $F$ and $G$ are also strongly associated.
Table 3 reports the univariate fraud estimates with corresponding confidence intervals. The confidence intervals are obtained with the parametric bootstrap method. When comparing the results of the LLRR and LLRRC models, it can be seen that the correction for cheating has a substantial effect on the univariate fraud estimates.

<table>
<thead>
<tr>
<th>Model</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>H0</td>
<td>0.07 (0.05, 0.11)</td>
<td>0.08 (0.06, 0.12)</td>
<td>0.11 (0.09, 0.15)</td>
<td>0.16 (0.12, 0.20)</td>
</tr>
<tr>
<td>H5</td>
<td>0.10 (0.07, 0.17)</td>
<td>0.11 (0.08, 0.17)</td>
<td>0.15 (0.11, 0.21)</td>
<td>0.25 (0.20, 0.32)</td>
</tr>
</tbody>
</table>

5 Robustness of the Model

In this section we evaluate the robustness of the cheating parameter estimate and the univariate prevalence estimates against non-zero $K$-factor interaction. We fitted the unconstrained LLRRC model to the data sets $n_{(−1)}$, $n_{(0)}$ and $n_{(1)}$, that were computed using the cheating and log-linear parameters of model H5 supplemented with a 4-factor interaction parameter $\lambda^{DEFG} \in \{-1, 0, 1\}$. The results are shown in Table 4. The columns labeled “True” show the true parameter values, and the columns labeled “Est.” show the estimates of the unconstrained LLRRC model. If $\lambda^{DEFG} = 0$ the estimates are unbiased, and otherwise the bias is relatively small.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$n_{(0)}^*$</th>
<th>$n_{(−1)}^*$</th>
<th>$n_{(1)}^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta$</td>
<td>.146</td>
<td>.146</td>
<td>.146</td>
</tr>
<tr>
<td>$\pi_1 (D)$</td>
<td>.104</td>
<td>.135</td>
<td>.094</td>
</tr>
<tr>
<td>$\pi_1 (E)$</td>
<td>.110</td>
<td>.151</td>
<td>.096</td>
</tr>
<tr>
<td>$\pi_1 (F)$</td>
<td>.150</td>
<td>.189</td>
<td>.137</td>
</tr>
<tr>
<td>$\pi_1 (G)$</td>
<td>.249</td>
<td>.539</td>
<td>.151</td>
</tr>
</tbody>
</table>

References


Bayesian inferences for object detection

Rozenn Dahyot

1 School of Computer Science and Statistics, Trinity College Dublin, Ireland, Rozenn.Dahyot@cs.tcd.ie

Moghaddam and Pentland (1997) first proposed to perform object detection by modelling the marginal probability density function (pdf) of high dimensional features of appearance. Based on gaussian hypotheses, their approach is however not robust to outliers that can occur in images due, for instance, to cluttered backgrounds or partial occlusions. In Dahyot et al. (2004), robustness has been improved by using better priors for the distribution of the errors encountered in the observations. Because the marginal pdf was not analytically available, the Maximum A Posteriori (MAP) pdf was used instead for detection. If both pdfs, marginal and MAP, are proportional under Gaussian assumptions (MacKay, 1995), and therefore perform equivalently, we aim in this paper to compare both densities (marginal and MAP) for object detection using the same robust priors.

Dimensionality reduction and intra class variability.

A collections of high dimensional features \( \{ x_k \}_{k=1,\ldots,K} \) are used to capture the variability of appearance in an object class \( O \). The \( N \)-dimensional vector \( x_k \) usually corresponds to colour or grey-level values of image pixels. It is assumed that the information contains in this data set can be captured by a variable \( \Theta \) of smaller dimension \( J << N \). The latent variable \( \Theta \) retains the main variabilities in the class \( O \) and is linked with the variable \( x \) by \( x = f(\Theta) + w_r \) where \( w_r \) is a reconstruction error. Several relations \( f \), linear or non-linear, have been proposed in the literature (Saul and Roweis, 2003). Here, a simple linear relation \( f \) has been learned from the training set using the Principal Component analysis (PCA) (Murase and Nayar, 1995, Moghaddam and Pentland, 1997, Dahyot et al., 2004).

Modelling Observations.

Features \( \{ x_k \}_{k=1,\ldots,K} \) used for training are assumed to be good representative of the object class, therefore clean from outliers. However, when comparing a new observation to the learned manifold, occurrences of gross errors due to cluttered background or partial occlusions, is a common issue. The observation \( y \) is then explicitly modelled as \( y = x + w^o = f(\Theta) + w \) where \( w^o \) represents the observation noise, and \( w \) is the sum of the reconstruction error and the observation noise.
Bayesian inferences.

A standard approach to obtain a measure of similarity between the observation $y$ and the class of interest is to compute the value of the marginal pdf $p_Y(y|O)$. Assuming the independence of variables $\Theta$ and $w$, the pdf can be computed by:

$$
\begin{align*}
p_Y(y|O) &= \int p_{Y,\Theta}(y, \Theta|O) \, d\Theta \\
&= \int p_{Y|\Theta}(y|\Theta, O) \cdot p_{\Theta}(\Theta|O) \, d\Theta \\
&= \int p_W(w|O) \cdot p_{\Theta}(\Theta|O) \, d\Theta
\end{align*}
$$

(1)

Another possible similarity measure corresponds to the value in $y$ of the MAP pdf (Jordan and Weiss, 2002, Dahyot et al., 2004):

$$
\begin{align*}
p_Y^{MAP}(y|O) &= \max_{\Theta} p_{Y,\Theta}(y, \Theta|O) = p_{Y,\Theta}(y, \hat{\Theta}_{MAP}|O)
\end{align*}
$$

(2)

The class $O$ is now characterised by a family of pdfs indexed by the latent variable $\Theta$. This approach consists in first selecting the best function that explained the observation $y$, and then computing its corresponding pdf value. The first step is performed in estimating $\hat{\Theta}_{MAP} = \arg \max_{\Theta} p_{Y,\Theta}(y, \Theta|O)$, and then computing $p_{Y,\Theta}(y, \hat{\Theta}_{MAP}|O)$.

Finally, now considering $\Theta$ as a parameter indexing the family of pdfs characterising the class $O$, the value in $y$ of the maximum likelihood (ML) pdf can also be computed as a similarity measure:

$$
\begin{align*}
p_Y^{ML}(y|O) &= \max_{\Theta} p_{Y|\Theta}(y|\Theta, O) = p_{Y|\Theta}(y|\hat{\Theta}_{ML}, O)
\end{align*}
$$

(3)

where $\hat{\Theta}_{ML} = \arg \max_{\Theta} p_{Y|\Theta}(y|\Theta, O)$.

Priors.

The integral in equation (1) has been solved analytically using normal assumptions (Tipping and Bishop, 1997, Moghaddam and Pentland, 1997). It is however difficult to solve it when expressions of $P_W(w|O)$ and $P_{C}(\Theta|O)$ are complex. Following Dahyot et al. (2004), the distribution of the error is modelled using hard redescender robust M-estimator function. This expression allows us to deal with gross error occurring in the observations.

The pdf of $\Theta$ is modelled in this experiment with the empirical distribution inferred using each training sample $x_k$ for which the latent variable $\Theta_k$ is estimated with $f$ in the training stage:

$$
p_{\Theta}(\Theta|O) = \frac{1}{K} \sum_{k=1}^{K} \delta(\|\Theta - \Theta_k\|)
$$

(4)

Under those assumptions, the marginal is computed by summing all probabilities computed for all possible $K$ latent variables $\Theta_k$. The MAP probability retains only the maximum which allows the possibility of speeding up
the exhaustive search. For the ML pdf, deterministic algorithms are used to compute for each observation $y$ the maximum likelihood estimate $\hat{\Theta}_{ML}$ (Dahyot et al., 2004). Other priors can also be used (Dahyot et al., 2004, Vik et al., 2003) with different robust estimation strategies (Dahyot et al. (2004), Vik et al., 2003, Dahyot and Wilson, 2005).

Detection: classification between $O$ and $\overline{O}$

The detection of objects is performed by first spanning the set of observations and computing their probabilities. Receiver Operating Characteristic (ROC) curves are used to assess our detectors, by computing the detection rate (number of detected occurrences of $O$ divided by the number of occurrences of $O$ present in the test set) w.r.t. the false alarm rate (number of detected occurrences of $\overline{O}$ divided by the number of occurrences of $\overline{O}$ present in the test set) over a test set of images of objects and non-objects ($\sim 500$ occurrences of $O$ and $\sim 60000$ occurrences of $\overline{O}$). Most samples of the objects in the test set present cluttered background and partial occlusions.

Figure 1 compares the ROC curves computed using all measures. In the test, the MAP pdf gives the best results, followed by the ML pdf and then the marginal pdf. The ML pdf performance is surpassed by the marginal at the top of the ROC curves. It is understood that in some difficult cases, when the observation shows too many outliers, the estimation of $\hat{\Theta}_{ML}$ is not accurate and, as a consequence, the ML pdf becomes less efficient than the marginal one. Concerning the computation time, the ML pdf, using a dedicated deterministic algorithm is the fastest followed by the MAP pdf and then the marginal.

Discussion.

If both pdfs, marginal and MAP, are proportional under gaussian assumptions (MacKay, 1995), this first experiment shows that better performances for detection can be expected when using the MAP under more realistic non-gaussian hypotheses. Both the MAP and ML pdfs perform better than the marginal. This experience gives better insights on the difference of the detection strategies presented in Moghaddam and Pentland (1997) and Dahyot et al. (2004), and shows the potential superiority of using the MAP pdf as a detector.

References


FIGURE 1. ROC curves: the MAP pdf (dash dot red, top), the ML pdf (continuous green, middle) and the marginal (long dash blue, bottom).

Department of Statistics, Ireland.


Statistics for network management

T. Dasu, J. Pelletier and J. Wright

Abstract: The field of statistics has undergone significant changes in the last decade in order to analyze data arising in new problem domains such as bioinformatics and internet traffic. Data sets in extant domains such as telecommunications have increased in size and complexity due to: new, sophisticated services and products; complex technological components; and ability to collect, monitor and access gigantic amounts of data. Network data epitomize many challenges of contemporary data sets such as rapid rates of accumulation, complex metadata and important application specific information e.g., the configuration data that describe the network topology. In this paper, we use network management analysis to illustrate the ways in which statistical methodology has adapted in order to understand and analyze these complex data sets and data streams, often by borrowing from other disciplines such as computer algorithms and artificial intelligence.

Keywords: Network management; nonparametric statistics; knowledge engineering.

1 Introduction

The last two decades have seen major revolutions in science and technology, with an accompanying explosion of data generated by new problem domains such as bioinformatics and internet security. Existing domains such as telecommunications and astronomy have grown exponentially with increasing technological sophistication, adding to the complexity and size of data, as well as the ever increasing speed of accumulation. Data are often accompanied by extensive metadata, context databases and knowledge bases that describe the application domain. Data sets are no longer simple row-by-column matrices of numbers or categories, but contain embedded images, audio/video, webpages, URLs and more. Statisticians need to incorporate these auxiliary data sources into their analysis in order to draw meaningful conclusions from the data.

In the past, statistics has relied on parametric methods based on well known distributions for the problems of estimation and inference. Data analysis used a model driven approach (e.g. logistic regression, proportional hazards) relying on asymptotic behavior of sampling distributions for studying errors and goodness-of-fit. These approaches, while analytically appealing, tend to be computationally intensive requiring multiple passes over
The underlying model and distributional assumptions are fairly strong, making them applicable in a limited set of scenarios. Nonparametric techniques exist, but are mainly univariate because they rely on rank ordering of data. The related discipline of data mining has focused on methods such as clustering, classification and association rules with an emphasis on performance and speed, rather than goodness of fit.

As the data sets become bigger, more complex and more heterogeneous, we need statistical methods that are fast, free of model and distributional assumptions and are easy to understand and interpret. These methods must also be able to assimilate complex metadata and domain knowledge. Other disciplines such as computer science and artificial intelligence offer technologies like space partitioning algorithms and expert systems that can augment statistical techniques.

In this paper, we use our experience with network management to illustrate the challenges of large, complex data and the ways in which we have adapted statistical techniques and combined them with other disciplines to analyze and mine these data sets.

2 Network Data

The sheer scale and intricacy exhibited by large networks is mind-boggling. An integrated view with the depth and breadth necessary to model, control, maintain, diagnose, and analyze large scale networks requires a wide spectrum of data. Systematically identifying and collecting essential data is no mean feat, but it pales besides the data mining issues raised - data integration, data interpretation, and data summarization. In addition, there are significant data quality issues that need to be addressed for consistent and reliable analyses, some of which are highly specific to network data, (Pelletier and Dasu, 2005). Furthermore, the data sources themselves are varied. For example, the defense of vital assets using tools like firewalls, intrusion detection systems (IDS), and vulnerability scanners (Cheswick, Bellovin and Rubin, 2003) create data sets of a description of the traffic that passes through them. Additionally, as seen in Fig 1, a typical network offers other opportunities to collect data about the state or status of network elements and nodes, how data is flowing through the network and data about the contents of the packets themselves. This collection of data streams and data sets can be grouped into four broad categories: configuration data, log data, traffic data, and alarms and exceptions.

Configuration data, in our context, refers to relatively static topological data and device information. Topological data, often represented using graph models, describes how the various network elements are interconnected. Device information contains the details of the available features, how each is configured, what services are active and how they are being used. Log data, generated by a device or an application, represents a running commentary on operational housekeeping, status, alarms and other
general information. This type of data tends to be voluminous, difficult to parse, and riddled with data quality issues. However, such data, while daunting, often contains the first indications of trouble and it should be analyzed carefully. A unique feature of the log data is the need for domain and context knowledge to facilitate data reduction and data interpretation. For example, an administrator might tag a power outage as innocuous if it is a part of a planned maintenance. Or, an apparent discontinuity in device logs may be due to a change in device name. Such context information needs to be recorded in knowledge bases and incorporated into the statistical analysis. While the first two categories define physical devices and their activities, traffic data (netflows) provides information on the flow of data packets through the network. For example, netflow data provides a data stream, which labels each packet with source IP, destination IP, flow duration, total bytes, source port, IP protocol and IP service type. Because traffic data is generated continuously, the challenge is accumulation and storage. Such data streams present unique challenges and are of particular interest (Muthukrishnan 2003). The fourth category deals with the contents of the packets moving across a network. For example, transmitted packets may contain worms or represent an intrusion attempt by hackers. Each packet is examined using signature matching rules defined in a knowledge base to identify potential viruses or intrusions and present the results in the form of alarm packets with attached error codes (Snort, 2004). Conversely, using computational statistical techniques as well as traditional statistical ones, we define network anomaly detection algorithms that detect departures from expected behavior.

3 Multi-pronged approach for multi-faceted data

Statistics plays a central role in identifying and modelling typical traffic patterns, and detecting departures from these expected patterns.
3.1 Data Quality, EDA

At the very outset, we applied data quality techniques and metrics to ensure that the data was certifiably accurate for running the analysis. In order to do this, we incorporated domain knowledge from experts into a knowledge base that was implemented in the preprocessing scripts. Ultimately, this process will be automated using an expert system. We then, used EDA (Exploratory Data Analysis) techniques to create summaries of longitudinal data to look for obvious anomalies – for instance, in Fig 3, we have shown three types of network and device events corresponding roughly to a two year period. The red events correspond to device A, the green ones to device B, and the background of grey events corresponds to events that are routine or are of no interest to us. The X-axis represents time, and the Y-axis is merely used to separate the devices of interest (device A and device B) from the rest of the network devices.

There are a couple of curious facts that emerge from this picture. First, there are two broad white gaps in the grey background of almost continuous occurrence of network events. The gaps correspond to periods during which there was a complete absence of data. The second gap corresponded to a known and scheduled outage, but the first broad gap could not be explained. While there is no way to recover data going back in time, we have put mechanisms in place that will alert us to missing data so that we can recover the data in a timely fashion.

The second oddity is the fact that device A and device B had logs for nonoverlapping time intervals, with similar event patterns. Our subject matter experts identified device B to a renamed, improved version of device A. We can see that in the reduction of events for device B, other than the equally spaced scheduled maintenance related events.

Longitudinal analysis of event patterns and correlating multiple types of event sequences is a challenging task, both in terms of synchronizing the data feeds as from multiple sources, as well as the complexity of the data. Our current research focuses on data collection algorithms that will address the synchronization issue, as well as using point process models to capture the statistical behavior of these event patterns.

3.2 Graph Models

A network is defined by the topology i.e. the interconnections between devices and networks. The topology plays an important role in studying traffic patterns, propagation of errors, performance and network security. Graph models are a natural choice to represent and study the network topology. We used simple criteria to study graphs of varying complexity. Fig 2 shows three patterns observed on our network. The center of each circle represents a network node (e.g. a server) and each radial line signifies traffic to another computer or IP address, denoted by a point on the
perimeter of the circle. The length of the radial represents the amount of traffic between the two. The first circle represents heavy traffic between two computers. The third circle corresponds to a central server communicating regularly with other computers on the same network. In contrast, the second circle shows highly suspicious activity. The fan-like pattern suggests that a single computer is sending substantial traffic to sequential IP addresses. This behavior is typical of an intruder searching for vulnerable computers by sequentially scanning an IP space.

3.3 Change Detection in Data Streams

Network traffic data streams consist of measurements of packets, payloads, packet loss and among other attributes. The streams are characterized by a rapid rate of accumulation, and typically, just a single opportunity to look at the data as it streams by. Some reduced, summarized versions of this data are stored but the raw data is typically too large to be stored in its entirety. An important task while analyzing data streams is detecting changes in distributions. In our past research, we proposed a nonparametric approach based on Kullback-Leibler distance for detecting shifts in distributions in data streams, in conjunction with efficient space partitioning algorithms and bootstrapping (Dasu, Krishnan, Venkatasubramanian and Yi, 2006). This approach was successfully used in determining stationarity of Internet measurements (Krishnamurthy, Madhyastha, Venkatasubramanian, 2005), and despite its generic nature and absence of prior knowledge, performed as well as a more domain specific algorithm. Our research (in progress) on correlating network traffic with network security events relies on this methodology.

In order to gain a complete understanding of the data, we incorporated varied data types including the hierarchical configuration data that describes the network; used computationally fast methods to deal with the issue of scale presented by the magnitude of the data; addressed the rapid rate of accumulation by focusing on summary based analysis; and used context information and domain knowledge repositories.
4 Conclusion

While we have described a network management application, the challenges posed by the data as well as the interdisciplinary approach needed to understand and analyze such data are typical of the complex datasets that we face today. The field of statistics has made headway in adapting to these changes in many ways, including a greater emphasis on computation and visualization. However, the nature of data has changed from the traditional data matrix, to include images, audio/video, complex semistructured metadata, domain specific knowledge embedded in rule bases and other data forms such as free text and tables from web pages. Statistics needs to evolve by adopting methodologies from multiple disciplines to seamlessly integrate and analyze such data.

References


FIGURE 3. Network Events.
Optimal design for linear mixed models

Legesse Kassa Debusho$^1$ and Linda M. Haines$^2$

$^1$ Department of Statistics and Operations Research, University of Limpopo, Pietersberg, South Africa, legessek@ul.ac.za
$^2$ Department of Statistical Sciences, University of Cape Town, Private Bag, Rondebosch 7700, South Africa, lhaines@stats.uct.ac.za

Abstract: This paper is concerned with developing explicit expressions for optimal designs, and specifically for $V$-optimal population designs, for the simple linear regression model with a random intercept term and with values of the explanatory variable taken from a set of equally spaced time points.

Keywords: Linear Regression; longitudinal data; population designs; $V$-optimality; $D$-optimality.

1 Introduction

The present study is broadly concerned with optimal experimental design for linear mixed models fitted to longitudinal data when the fixed effects parameters are of particular interest. The essential problem is that of choosing the numbers of individuals to be allocated to various groups or cohorts and of choosing the times for taking measurements on the individuals within each group. Earlier studies are reviewed in, for example, Ouwens, Tan and Berger (2002) and provide approaches to design construction which are somewhat diverse. The immediate aim here is to develop some underlying and cohesive theory and, in particular, to establish algebraic results in the context of the simple linear regression model with a random intercept.

2 An Example

Duchateau, Janssen and Rowlands (1998, p. 13) reported the results of an experiment to study the susceptibility of cattle to trypanosomosis or animal sleeping sickness. The data set included a series of measurements on the changes in packed cell volume (PCV), which is the percentage of the volume of the blood serum taken up by the red blood cells, following experimental infection with trypanosomes for six animals from the N’Dama breed of cattle. The animals were sampled over a series of fourteen different, unequally spaced days coded as 0, 2, 4, 7, 9, 14, 17, 18, 21, 23, 25, 29, 31 and 35 and a simple linear regression model with a random intercept term
provided an excellent fit to the data. Specifically, the model for \( y_{ij} \), the \( j \)th observation on the \( i \)th individual at time point \( t_{ij} \), is given by

\[
y_{ij} = \beta_0 + \beta_1 t_{ij} + b_i + e_{ij}, \quad j = 1, \ldots, 14, \quad i = 1, \ldots, 6, \tag{1}
\]

where the term \( b_i \) is a random effect associated with the \( i \)th individual and is distributed as \( \mathcal{N}(0, \sigma^2_b) \), the term \( e_{ij} \) represents a measurement error distributed as \( \mathcal{N}(0, \sigma^2_e) \), and \( b_i \) and \( e_{ij} \) are independent both within and between individuals. The intercept \( \beta_0 \) and the slope \( \beta_1 \) are fixed effects with MLE’s given by \( \hat{\beta}_0 = 35.077 \) and \( \hat{\beta}_1 = -0.276 \) respectively and the parameters \( \sigma^2_b \) and \( \sigma^2_e \) comprise the variance components with REML estimates of \( \hat{\sigma}_b^2 = 4.181 \) and \( \hat{\sigma}_e^2 = 3.595 \) respectively.

The problem addressed here is that of redesigning the experiment in order to estimate the model parameters \( \boldsymbol{\beta} = (\beta_0, \beta_1) \), or some function of them, as precisely as possible. For example, should the researcher take 2 measurements on days 0 and 35 on 42 animals, or should the researcher rather take an allocation of observations to days based on say 7 or 12 animals?

### 3 V-Optimal Designs

#### 3.1 Theory

Consider model (1) in a more general form for \( K \) individuals, \( d_i \) observations taken on the \( i \)th individual and time points for each individual taken without replacement from the set \( \{0, 1, \ldots, k\} \). Then an individual design can be specified by a vector of time points \( t \) and a design over all individuals, termed a population design, by an allocation of \( n_j \) individuals to the distinct designs \( t_j \) for \( j = 1, \ldots, r \). In fact, within the context of optimal design theory, it is convenient to introduce an approximate population design based on \( r \) distinct individual designs as a probability measure \( \xi \) on a space of designs which assigns a weight \( w_j \) to the \( j \)th individual design \( t_j \) on a per observation basis, that is

\[
\xi = \left\{ \frac{t_1}{w_1}, \ldots, \frac{t_r}{w_r} \right\} \quad \text{where} \quad 0 < w_j < 1 \quad \text{and} \quad \sum_{j=1}^{r} w_j = 1.
\]

Then the standardized information matrix for \( \boldsymbol{\beta} \) at the approximate population design \( \xi \) is given by \( M_{\beta}(\xi) = \sum_{j=1}^{r} w_j M_{\beta}(t_j) \), where \( M_{\beta}(t_j) \) is the information matrix at the design \( t_j \) evaluated on a per point basis. Note that \( M_{\beta}(\xi) \) depends only on the intra-class correlation coefficient \( \gamma = \sigma^2_b / \sigma^2_e \) and not on the fixed effects \( \boldsymbol{\beta} \) or, up to to a multiplying constant, on \( \sigma^2_e \).

Suppose now that interest centres on estimating the mean marginal responses at the given vector of time points \( t_g = (0, 1, \ldots, k) \) as precisely as possible. Then the \( V \)-optimality criterion, which is proportional to the average of the asymptotic variances of the estimators of the population mean
responses at $t_g$, can be formulated as

$$
\Psi_V(\xi) = \text{tr}\{X_g M^{-1}_g(\xi) X^T_g\} = \text{tr}\{M^{-1}_g(\xi) X^T_g X_g\}
$$

where $X_g = [1 \ t_g]$ and a $V$-optimal design can in turn be defined as a design for which $\Psi_V(\xi)$ is minimized over the space of all possible population designs. Two key results can be derived by using arguments based on the geometry of the design space and by invoking an Equivalence Theorem related to theorems for population designs given in Cheng (1995) and Mentré, Mallet and Baccar (1997) and these are summarized as follows.

Result 1: The $V$-optimal population design for the mean responses at $t_g$ over the space of all $d$-point designs for all values of the parameter $\gamma \geq 0$ is given by

$$
\xi^*_V = \{(0, 1, \ldots, d - \frac{d}{2} - 1, k - \frac{d}{2} + 1, \ldots, k - 1, k)\}
$$

for $d$ even and for $d = k + 1$ with $d$ odd and by

$$
\xi^*_V = \{(0, 1, \ldots, d - \frac{d}{2} - 1, k - \frac{d}{2}, \ldots, k) \ (0, 1, \ldots, d - \frac{d}{2}, k - \frac{d}{2} + 1, \ldots, k)\}
$$

otherwise.

Result 2: The $V$-optimal population design for the mean responses at $t_g$ over the space of all possible individual designs is given, for $0 < \gamma \leq \frac{k - 1}{k + 2}$, by

$$
\xi^*_V = \{(0) \ (k)\}
$$

and, for $\gamma > \frac{k - 1}{k + 2}$, by

$$
\xi^*_V = \{(0) \ (k) \ (0, k)\}
$$

where

$$
w = \frac{(\gamma + 1) \{k (3 \gamma + 2) + 1 - \sqrt{3k(k + 2)(2 \gamma + 1)}\}}{2 \gamma (3k \gamma + k - 1)}.
$$

3.2 Example

Consider now redesigning the experiment on N’Dama cattle. Assume that only 84 observations can be taken, as in the original experiment, and that $\gamma$ is given by its REML estimate, $\hat{\gamma} = 1.163$. It follows immediately from Result 1 that if 6 observations are to be taken on 14 animals then for $V$-optimality these should be on days 0, 1, 2, 33, 34 and 35 for all animals, whereas if 7 observations are to be taken on 12 animals, then these
should be on days 0, 1, 2, 3, 33, 34 and 35 for 6 of the animals and on days 0, 1, 2, 32, 33, 34 and 35 for the remaining 6 animals. Furthermore, since the assumed value of $\gamma$ is greater than 1 and thus greater than $\frac{k - 1}{k + 1} = \frac{17}{18}$, it follows from Result 2 that the best $V$-optimal population design over the set of all possible such designs puts weights, on a per point basis, of 0.4732 on the designs with single measurements on day 0 and day 35 and of 0.0536 on the two-point design with support at days 0 and 35. An exact design, on a per animal basis, can be constructed directly from this approximate design and comprises 41 animals allocated to day 0, 41 to day 35 and 2 to the two-point design (0, 35). An immediate problem with this latter design is that the single point individual designs do not provide information on the variance components.

4 Extensions

4.1 $D$-optimal Designs

Consider again the model described in Section 3.1 and suppose now that interest centres on the precise estimation of the parameters $\beta_0$ and $\beta_1$. Then the $D$-optimality criterion defined by

$$\Psi_D(\xi) = \ln |M_\beta(\xi)|$$

and thus relating to the generalized variance of the parameter estimates is appropriate. Key results relating to $D$-optimal population designs can immediately be derived in a manner analogous to that invoked for $V$-optimality. Specifically Result 1, which provides an explicit form for the optimal designs over the space of all possible $d$-point designs, also holds for $D$-optimality. However the population design which is $D$-optimal over all possible designs is simply the two-point design comprising the extreme points 0 and $k$ and does not coincide completely with the $V$-optimal population designs specified in Result 2. This $D$-optimal population design is intuitively appealing and permits estimation of the variance components but is, in an academic sense, less interesting than its $V$-optimal counterpart.

4.2 Designs for Random Coefficient Models

Suppose now that a simple linear regression model with a random intercept and, additionally, a random slope provides an appropriate fit to a given data set. Then the model can be formulated as the random coefficient model

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) t_{ij} + e_{ij}, \quad j = 1, \ldots, d, \quad i = 1, \ldots, K,$$

(2)

where the terms $b_{0i}$ and $b_{1i}$ are random effects associated with the intercept and the slope for the $i$th individual respectively and are normally distributed, independently of the measurement error, with zero means and with
variance matrix $G$. A crucial feature of such models is that a change in scale and location of the explanatory variables induces a change in the matrix $G$. This has an immediate impact on design. For example, consider model (2) with 5 equally spaced time points available, with a random slope specified by $b_{1i} \sim N(0, \sigma_1^2)$, with a fixed intercept, and with a variance ratio $\sigma_1^2/\sigma_2^2$ of 0.1. Then for non-replicated time points taken from the set $\{0, 1, 2, 3, 4\}$, the $D$-optimal population design for the precise estimation of the fixed effects $\beta$ allocates all individuals to the design $(0, 1, 4)$ whereas for time points from the set $\{-2, -1, 0, 1, 2\}$, the corresponding design allocates half of the individuals to the design $(-2, -1, 2)$ and half to $(-2, 1, 2)$. In fact the $D$-optimal population design for the coding $\{0, 1, 2, 3, 4\}$ is optimal for a random coefficient model with singular variance matrix $G = \sigma_1^2 \begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix}$ for the coding $\{-2, -1, 0, 1, 2\}$. Clearly a random coefficient model, together with the coding of the explanatory variables, must be specified with great care and in addition optimal population designs must be correctly associated with the underlying model.

5 Discussion

There are a number of extensions to the work reported here. In particular other criteria relating to the fixed effects can be explored and the scope of interest can be extended to the precise estimation of the variance components.

Acknowledgments: The authors would like to thank the University of KwaZulu-Natal, the University of Limpopo, the University of Cape Town and the National Research Foundation, South Africa, for financial support and Luc Duchateau and Tim O’Brien for insightful comments.

References


Mixed models, array methods and multidimensional density estimation

Maria Durbán¹, Iain Currie² and Paul H.C. Eilers³

¹ Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, Madrid, Spain, mdurban@est-econ.uc3m.es
² Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, EH14 4AS, Scotland
³ Department of Medical Statistics, Leiden University Medical Center, 2300 RC Leiden, The Netherlands

Abstract: We consider multidimensional density estimation. Eilers et al. (2006) used multidimensional P-splines together with a new fast and compact algorithm to smooth multidimensional histograms. Currie et al. (2006) emphasized the array nature of the method and demonstrated that a mixed model could also be fitted into this array structure. In this paper we combine these ideas and show how array methods and mixed models give rise to an efficient method of estimating multidimensional densities.

Keywords: Array methods; density estimation; mixed models; P-splines.

1 Array methods

In the companion papers, Eilers & Marx (2006) and Lambert & Eilers (2006) describe efficient methods for multidimensional density smoothing using P-splines; in these papers, the smoothing parameters are chosen by maximizing some information criterion, such as AIC, or by Bayesian methods respectively. Our paper tackles the same problem and uses the same computational ideas but differs from these papers in two important respects. First, we adopt a compact array notation; this has the advantage of clarifying the close connection between some seemingly unrelated computations. Second, we choose the amount of smoothing by expressing the model as a generalized linear mixed model; the smoothing parameters are then chosen by maximizing the residual log likelihood (REML) rather than by maximizing some information criterion, such as AIC.

For simplicity we will describe the 2-dimensional case and comment on the general case in our concluding remarks. We suppose that we have count data arranged in a 2-dimensional array or matrix $Y$, $n_1 \times n_2$. The rows of $Y$ are indexed by $x_1 = (1, \ldots, n_1)'$ and the columns by $x_2 = (1, \ldots, n_2)'$. Let $B_1 = B_1(x_1)$, $n_1 \times c_1$, be a marginal model matrix of equally spaced
$B$-splines for data $x_1$; similarly, let $B_2 = B_2(x_2)$, $n_2 \times c_2$, be a marginal model matrix for data $x_2$. The Kronecker or tensor product, $B = B_2 \otimes B_1$, gives a suitable basis for 2-dimensional smoothing. If we assume that the individual counts follow a Poisson distribution then we have defined a generalized linear model or GLM with data vector $y = \text{vec} Y$ and model matrix $B$. Estimates of the regression coefficients are then obtained by repeated evaluation of the scoring algorithm

$$B'\hat{W}_\delta B\hat{\theta} = B'\hat{W}_\delta \hat{z}$$  \hspace{1cm} (1)$$

where $\hat{z}$ is the working variable, $\hat{W}_\delta$ is a diagonal weight matrix and $\hat{\theta}$, $c_1c_2 \times 1$, is the vector of regression coefficients. If we have chosen a large number of basis functions in $B_1$ and $B_2$ then the resulting surface will be undersmoothed without penalization. Let $\Theta$, $c_1 \times c_2$, be the matrix of regression coefficients where $\text{vec} \Theta = \theta$. Penalization of the rows and columns of $\Theta$ (marginal penalization) gives

$$P = \lambda_1 I_{c_2} \otimes D_1' D_1 + \lambda_2 D_2' D_2 \otimes I_{c_1}$$  \hspace{1cm} (2)$$

as the penalty matrix where $D_1$ and $\lambda_1$, and $D_2$ and $\lambda_2$ are difference matrices of orders $q_1$ and $q_2$, and smoothing parameters on the columns and rows of $\Theta$ respectively. Algorithm (1) becomes

$$(B'\hat{W}_\delta B + P)\hat{\theta} = B'\hat{W}_\delta \hat{z},$$  \hspace{1cm} (3)$$

the penalized scoring algorithm, and appropriate values of $\lambda_1$ and $\lambda_2$ yield the smoothed surface.

Algorithms (1) and (3) are perfectly general but we now take advantage of the array structure of the data and the Kronecker product structure of the model matrix and show that (1) and (3) can be evaluated without the evaluation of the full model matrix and at hugely increased speed. We require three quantities: the two linear functions $B\theta$ and $B'\hat{W}_\delta \hat{z}$ and the weighted inner product $B'\hat{W}_\delta B$. The linear functions are dealt with by a well-known matrix identity

$$(B_2 \otimes B_1)\theta, n_1n_2 \times 1 \equiv B_1\Theta B_2', n_1 \times n_2;$$  \hspace{1cm} (4)$$

here $\equiv$ means that both sides have the same elements, although their dimensions, as indicated, are different.

To compute the weighted inner product we use the row tensor function of two matrices, $X_1$, $n \times c_1$, and $X_2$, $n \times c_2$,

$$G(X_1, X_2) = (X_1 \otimes 1'_{c_2}) * (1'_{c_1} \otimes X_2)$$  \hspace{1cm} (5)$$

where $1_c$ is a vector of 1’s of length $c$ and $*$ indicates element-by-element multiplication. Then $B'\hat{W}_\delta B$ is computed with

$$(B_2 \otimes B_1)'\hat{W}_\delta (B_2 \otimes B_1), c_1c_2 \times c_1c_2 \equiv G(B_1, B_1)'WG(B_2, B_2), c_1^2 \times c_2^2$$  \hspace{1cm} (6)$$
where $W$ is the $n_1 \times n_2$ matrix of weights such that $\text{vec} W = \text{diag} W$. This formula has the same form as (4) and the right-hand side of (6) enables the efficient evaluation of the inner product without the evaluation of $B$. There is a small overhead required to reorganize the right-hand side of (6) into the left-hand side; see Eilers et al. (2006) or Currie et al. (2006) for details.

2 A generalized linear mixed model approach

Mixed models and, more generally, generalized linear mixed models (GLMM) are popular and widely available so it is of interest to describe how array methods may be applied with this approach. The GLMM approach consists of two stages: (a) reparameterizing the linear predictor and (b) estimating the model parameters. In section 2.1 we give a mixed model representation of a penalized regression with normal errors; in section 2.2 we give the estimates of the mixed model parameters and extend the normal model to a GLMM; and in 2.3 we show how the necessary computations may be accomplished with the array arithmetic of (4) and (6). We describe these three stages in turn.

2.1 Reparameterizing the linear predictor

There are several approaches to the formulation of splines, and in particular of P-splines, as mixed models. These approaches differ mainly in the bases used for regression and in the penalty, and so yield mixed models with different fixed and random parts (although the fitted models are equivalent). We wish to reparameterize

$$y = B\theta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I) \quad (7)$$

where $B = B_2 \otimes B_1$ and smoothness is imposed via the penalty matrix $P$ in (2). The aim is to look for a new basis which allows the representation of (7) with its associated penalty as a mixed model

$$y = X\beta + Z\alpha + \epsilon, \quad \alpha \sim \mathcal{N}(0, G), \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I) \quad (8)$$

where $G$ is a diagonal matrix which depends on $\lambda_1$ and $\lambda_2$. Furthermore, we have the additional requirement that we want to be able to compute the resulting mixed model with the array algorithms (4) and (6).

Our method uses the singular value decomposition (SVD) of the penalty $P$. We take the SVD of $D' D$ into $U_1 \Sigma_1 U_1'$ and partition the matrix $U_1 = [U_{1s} : U_{1n}]$, where $U_{1s}$ corresponds to the non-zero eigenvalues and $U_{1n}$ to the zero eigenvalues. In this paper we assume a second order penalty, $q_1 = q_2 = 2$, and so $\Sigma_1$ has two zero eigenvalues and $U_{1n}$ has two...
Multidimensional density estimation

columns. Let $\Sigma_{1s}$ contain the positive eigenvalues of $\Sigma_1$. We have similar
definitions for $U_{2s}$, $U_{2n}$ and $\Sigma_{2s}$ arising from the SVD of $D_{2}^{1}D_{2}^{\prime}$. Let

$$T = (U_{2n} \otimes U_{1n} : U_{2s} \otimes U_{1s} : U_{2s} \otimes U_{1s}) .$$

(9)
The model matrices $X$ and $Z$ for the fixed and random parts respectively
are defined by $BT = [X : Z]$ where

$$X = B(U_{2n} \otimes U_{1n}) = B_{2}U_{2n} \otimes B_{1}U_{1n} = X_{2} \otimes X_{1}$$

(10)
and

$$Z = B(U_{2s} \otimes U_{1n} : U_{2n} \otimes U_{1s} : U_{2s} \otimes U_{1s})$$

(11)

$$= (B_{2}U_{2s} \otimes B_{1}U_{1n} : B_{2}U_{2n} \otimes B_{1}U_{1s} : B_{2}U_{2s} \otimes B_{1}U_{1s})$$

(12)

$$= (Z_{2} \otimes X_{1} : X_{2} \otimes Z_{1} : Z_{2} \otimes Z_{1}) .$$

(13)
Now $T$ is orthogonal so we may reparameterize as $B\theta = X\beta + Z\alpha$ where

$$\beta = (U_{2n} \otimes U_{1n})' \theta$$

(14)
$$\alpha = (U_{2s} \otimes U_{1n} : U_{2n} \otimes U_{1s} : U_{2s} \otimes U_{1s})' \theta .$$

(15)
Lastly we consider the penalty. The penalty term $\theta' P \theta = \omega' T' PT \omega$ where

$\omega' = (\beta', \alpha')$. Now $P(U_{2n} \otimes U_{1n}) = 0$ so there is no penalty on the fixed
part and the penalty becomes $\alpha' F \alpha$ for some $F$. Further it is straightforward to show that $F$ is the diagonal matrix

$$F = \begin{pmatrix}
    \lambda_{2} \Sigma_{2s} \otimes I_{q_{1}} & \lambda_{1} I_{q_{2}} \otimes \Sigma_{1s} & \\
    \lambda_{1} I_{q_{2}} \otimes \Sigma_{1s} & \lambda_{2} \Sigma_{2s} \otimes I_{c_{1} - q_{1}}
\end{pmatrix} .$$

(16)
With the new basis and new penalty the penalized sum of squares

$$S(\theta; y, \lambda_{1}, \lambda_{2}) = (y - B\theta)'(y - B\theta) + \theta' P \theta$$

(17)
transforms into

$$S(\alpha, \beta; y, \lambda_{1}, \lambda_{2}) = (y - X\beta - Z\alpha)'(y - X\beta - Z\alpha) + \alpha' F \alpha .$$

(18)
It follows that the estimates of $\alpha$ and $\beta$ satisfy the mixed model equations
for the mixed model (8) with $X$ given by (10), $Z$ given by (13) and $G = \sigma^{2} F^{-1}$; see Searle et al. (1992, p276).

2.2 Estimation

Estimates of $\beta$ and $\alpha$ follow from standard mixed model theory

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$$

(19)
$$\hat{\alpha} = GZ'V^{-1}(y - X\hat{\beta})$$

(20)
where \( V = \sigma^2 I + ZGZ' \); smoothing parameters may be selected by maximizing the residual log likelihood (REML) \( \ell(\lambda_1, \lambda_2) \)

\[
-\frac{1}{2} \log |V| - \frac{1}{2} \log |X'V^{-1}X| - \frac{1}{2} y'(V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1})y.
\]

(21)

We now iterate between (19) and (20) on the one hand and (21) on the other.

To extend the analysis to a GLMM we use the penalized quasi-likelihood (PQL) approach of Breslow & Clayton (1993). The estimates given by PQL correspond to the estimates in a weighted linear Gaussian model with error variance \( \sigma^2 = 1 \). The matrix \( V \) in (19), (20) and (21) is given by

\[
V = W_\delta^{-1} + ZGZ'
\]

(22)

where the weights \( W_\delta \) are given by \( \text{diag}(\exp(X\beta + Z\alpha)) \) in the Poisson case. This formulation makes it possible to fit the model with the functions \( \text{lme( )} \) in Splus and R. The R function \( \text{glmm.PQL( )} \) may also be adapted to fit the model. In one dimension this approach is satisfactory but where we need to fit a surface this approach runs into computational difficulties.

2.3 Evaluation

We indicate how the Kronecker product structure of \( X \) and \( Z \) allows the array algorithms (4) and (6) to be applied to the evaluation of (19), (20) and (21). The solution of (19), (20) and (21) appears to involve the inversion of the \( n_1n_2 \times n_1n_2 \) matrix \( V \). However we may show that

\[
V^{-1} = W_\delta - W_\delta Z(G^{-1} + Z'W_\delta Z)^{-1}Z'W_\delta
\]

(23)

and

\[
|V| = |W_\delta|^{-1}|G||G^{-1} + Z'W_\delta Z|.
\]

(24)

and we need only invert a \( c_1c_2 \times c_1c_2 \) matrix in both cases; see Searle et al. (1992, p453). Furthermore, it is easy to check that the constituents of (19), (20) and (21), such as \( Z'W_\delta Z \), \( Z'W_\delta y \), etc., may be assembled with the algorithms (4) and (6). The computations avoid the calculation of the potentially large matrices \( X \) and in particular \( Z \), and are performed at greatly enhanced speed.

3 Examples

We discuss two examples: the first example is a simulation which is designed to show how the fixed and random parts can be rearranged into additive and non-additive components, and the second is an analysis of the “Old Faithful” geyser data.
3.1 A simulation exercise

Figure 1 shows data simulated on a $30 \times 20$ grid. We take $B_1$, $30 \times 13$, and $B_2$, $20 \times 10$ which gives a full regression matrix $B$, $600 \times 130$. Now $X_1 = B_1 U_{1n}$ where $U_{1n}$ consists of the null eigenvectors of $D_1'D_1$ and so $X_1$ (with a second order penalty) can also be taken as $(1_{n_1} : x_1)$; similarly, we take $X_2 = (1_{n_2} : x_2)$. We now expand $X$ and $Z$ in (10) and (13) as follows

$$X = (1 : x_2 \otimes 1_{n_1} : 1_{n_2} \otimes x_1 : x_2 \otimes x_1)$$

$$Z = (Z_2 \otimes 1_{n_1} : Z_2 \otimes x_1 : 1_{n_2} \otimes Z_1 : x_2 \otimes Z_1 : Z_2 \otimes Z_1).$$

This partition facilitates two things: first, we can write the fitted surface as an overall mean plus the sum of three terms, a term for $x_2$ (second block of $X$ and first block of $Z$), a term for $x_1$ (third block of $X$ and third block of $Z$) and an interaction term which depends on both covariates (the sum of the four remaining terms). The individual and interaction components and the fitted surface are also shown in Figure 1. Second, we can fit submodels such as the model consisting of the two additive terms only. This hierarchical approach can lead to model selection and model simplification.
3.2 Old Faithful

The “Old Faithful” geyser data set consists of 272 pairs of eruption duration times and waiting times till the next eruption of the “Old Faithful” geyser in Yellowstone National Park. To apply the present method we first construct a two-dimensional histogram over a fine grid; we used $100 \times 100$ in this example. Note that the great majority of the histogram bins have a zero count. We take $B$-spline bases $B_1$ and $B_2$ each of rank 13 which gives a full model matrix $B$ of size $10^4 \times 169$. A conventional regression approach would be quite challenging but with our array methods, selection of the smoothing parameters and estimation of the smoothed density took less than 40 seconds for both isotropic and anisotropic smoothing. Figure 2 shows the density contour plots for both isotropic and anisotropic smoothing.

![Density Contour Plots](image)

**FIGURE 2.** Estimated density with the isotropic model (left) and the anisotropic model (right)

We can compare the isotropic and anisotropic fits by using the Akaike Information Criterion (AIC). For the isotropic fit the AIC value was 1342.68 with effective dimension (ED) of 26.27 (as measured by the trace of the hat matrix); for the anisotropic fit the AIC value was 1334.69 with ED of
We conclude that the anisotropic model is necessary for a proper description of the 2-dimensional density.

4 Conclusions

Mixed models give a fast and compact method of 2-dimensional density smoothing. The method generalizes to higher dimensions where the advantages of array methods in terms of storage and speed are even greater. The necessary functions to generalize (4) and (6) to higher dimensions are given in Currie et al. (2006). In density estimation the advantages of array methods are compelling whether one favours smoothing parameter selection with an information criterion, with a Bayesian method or, as in the present case, with a mixed model representation.

This paper has developed a new basis from the original $B$-spline basis. The new basis (a) allows mixed model methods to be used to fit a $d$-dimensional surface, (b) enables the fitted surface to be expressed as a sum of additive and interaction terms and (c) clarifies the role of the penalty (or penalties) by relating the penalty to the additive terms in the model. These remarks apply more widely since the new basis can be used when either a Bayesian or an information criterion method is used for model selection: the benefits of the decomposition of the surface and the clarification of the penalty apply equally here.

References


Multidimensional density smoothing with P-splines

Paul H.C. Eilers, Brian D. Marx

1 Department of Medical Statistics, Leiden University Medical Center, 2300 RC, Leiden, The Netherlands, p.eilers@lumc.nl
2 Department of Experimental Statistics, Louisiana State University, Baton Rouge, Louisiana 70803, USA, bmarx@lsu.edu

Abstract: We propose a simple and effective multidimensional density estimator. Our approach is essentially penalized Poisson regression using a rich tensor product B-spline basis, where the Poisson counts come from processing the data into a multidimensional histogram, often consisting of thousands of bins. The penalty enforces smoothness of the B-spline coefficients, specifically within the rows, columns, layers—depending on the dimension. In this paper, we focus on how a one-dimensional P-spline density estimator can be extended to two-dimensions, and beyond. In higher dimensions we provide a hint on how efficient grid algorithms can be implemented using array regression. Our approach optimizes the penalty weight parameter(s) using information criteria, specifically AIC. Two examples illustrate our method in two-dimensions.

Keywords: AIC; effective dimension; tensor product.

1 Introduction

Density estimation is a core activity in data exploration and applied statistics. Surprisingly, many statisticians approach it in a quite unsophisticated way: they are happy with off-the-shelf kernel smoothers, as can be seen in many papers on computational Bayesian modeling. An attractive alternative is (penalized) likelihood smoothing, modeling the logarithm of the density with splines. Kooperberg and Stone (1991) use adaptive-knot B-splines, while Eilers and Marx (1996) combine fixed-knot B-splines with a roughness penalty (P-splines), simplifying the scheme of O’Sullivan (1988).

In the P-spline approach, the raw data are processed and reduced to a histogram, then density estimation is achieved through (penalized) Poisson regression. This is attractive for high-volume one-dimensional data, because one deals with, e.g., 100 or 200 histogram bins instead of many thousands of raw observations. In two or more dimensions the histogram approach is even more attractive, when combined with a recently developed algorithm for fast weighted smoothing on grids (Eilers et al., 2006).
A critical issue is optimization of the amount of smoothing. We use Akaike’s Information Criterion (AIC). It is easy to compute, because an effective model dimension can be defined, and it can be computed with relatively little effort.

2 P-spline overview: one-dimensional densities

Let \((y_i, u_i)\) denote the Poisson count and bin midpoint pairs from a (narrowly binned) histogram, \(i = 1, \ldots, n\). The vector of counts is denoted \(y\). We model the expected values of the counts as

\[
\mu_i = E(y_i) = \exp\left(\sum_{j=1}^{c} b_j(u_i) \theta_j\right),
\]

or in matrix terms \(\mu = B \theta\), where \(B = [b_j(u_i)]\) is a \((n \times c)\) B-spline basis built along the indexing axis \(u\) of the density. A rich basis is used (\(c\) sufficiently large) and the knots for the basis are equally-spaced. Apart from a constant, the Poisson log-likelihood is

\[
l = \sum_{i=1}^{n} \log(\mu_i y_i e^{-\mu_i}) = \sum_{i=1}^{n} (y_i \log \mu_i - \mu_i).
\]

A penalty on the \(d\)-th differences of \(\theta\) is subtracted from the log-likelihood, to tune smoothness. The number of basis functions in \(B\) is chosen large, to give ample flexibility. The penalized log-likelihood, \(l^*\), then is

\[
l^* = l - \lambda D' D \theta.
\]

Setting the gradient of (3) equal to zero gives

\[
B'(y - \mu) = \lambda D' D \theta,
\]

where \(D\) is a matrix of contrasts such that \(D\theta = \Delta^d \theta\). Linearization of (4) leads to

\[
(B'\hat{W}B + \lambda D'D)\theta = B'\hat{W}\hat{z},
\]

where \(z = \eta + \hat{W}^{-1}(y - \mu)\) is the working variable and \(\eta = B\theta\). The matrix \(W = \text{diag}(\mu)\) and \(\hat{\theta}, \hat{\mu}\) are approximations to the solution of the (5), iterated until convergence. One easily recognizes the familiar equations for fitting a GLM, modified by the term \(\lambda D'D\) that stems from the penalty. At convergence, one can interpret (5) as linear smoothing of the working variable \(z\). We have

\[
\hat{z} = B\hat{\theta} = B(B'\hat{W}B + \lambda D'D)^{-1}B'\hat{W}z = Hz,
\]
where $H$ is the smoothing or “hat” matrix. The effective dimension of the smoother is approximated by $\text{race}(H)$. We use it to define Akaike’s information criterion (AIC) as

$$AIC = 2 \sum_{i=1}^{n} y_i \log(y_i/\hat{\mu}_i) + 2\text{trace}(H).$$

(7)

Essentially this is all old hat. A more detailed description of density smoothing with P-splines is given by Eilers and Marx (1996). We use the presentation above as an impetus to move density smoothing with use tensor product P-splines in two or more dimensions.

3 Two dimensional P-spline densities

Let $Y_{ih}$ be counts in a two-dimensional (narrowly binned) $n_1 \times n_2$ histogram, forming the array $Y$ of Poisson counts with expectation $M$. The bin midpoints are now indexed by the $n_1 n_2$ pairs $(u_1 \otimes 1_{n_2}, 1_{n_1} \otimes u_2)$, where $u_1$ ($u_2$) are the midpoint locations along the first (second) dimension. We model the expected values, now using tensor product B-splines on a rich $c_1 \times c_2$ rectangular grid (equally-spaced along each dimension), as

$$\mu_{ih} = E(Y_{ih}) = \exp\left(\sum_{j=1}^{c_1} \sum_{k=1}^{c_2} b_j(u_{1i}) b_k(u_{2h}) \theta_{jk}\right).$$

(8)

The $n_1 n_2$-vectorized form of the Poisson counts and corresponding expected values are denoted by $y = \text{vec}(Y)$ and $\mu = \text{vec}(M)$, respectively. Ignoring penalties for the moment, (8) is also a GLM, and can be rewritten as

$$\mu = E(y) = B_2 \otimes B_1 \theta = B\theta.$$

(9)

We now have the $c_1 \times c_2$ array of coefficients $\Theta$, with vectorized form $\theta = \text{vec}(\Theta)$. The matrix $B$, with $n_1 n_2$ rows and $c_1 c_2$ columns, collects the tensor product basis functions.

As outlined in Durbán et al. (2004), the penalty can be written as

$$\theta' P \theta = \theta'(\lambda_1 I_{c_2} \otimes D_1' D_1 + \lambda_2 D_2' D_2 \otimes I_{c_1}) \theta$$

$$= \lambda_1 ||D_1 \Theta||_F^2 + \lambda_2 ||D_2 \Theta||_F^2,$$

(10)

where $||\cdot||_F^2$ indicates the quadratic Frobenius norm, the sum of the squares of all elements of a matrix, and $D_1(D_2)$ is the proper contrast matrix for the rows (columns) of $\Theta$. Such a penalty enforces smoothness or tensor product B-spline coefficients within any one row or column, but breaks linkage in penalization across rows or columns. As seen in (10), the weight and order of the penalty is the same from row to row, the same from column to column, but can vary from rows to columns. In this scheme, with $\eta = B\theta$, the equivalent of (5) becomes

$$(B' W B + P) \theta = B' W (\eta + W^{-1}(y - \hat{\mu})).$$

(11)
3.1 Optimizing the penalty

To optimize the amount of smoothing, AIC can be computed similarly to (7). Upon convergence AIC is now computed as

$$AIC = 2 \sum_{i=1}^{n_1} \sum_{h=1}^{n_2} y_{ih} \log(y_{ih}/\hat{\mu}_{ih}) + 2\text{trace}\{B'\hat{W}B(B'\hat{W}B + P)^{-1}\}, \quad (12)$$

borrowing the result that trace is invariant under cyclical permutation of matrices. One varies the logarithms of $\lambda_1$ and $\lambda_2$ on a grid, searching for the minimum. In many cases, or for first exploration, isotropic smoothing ($\lambda_1 = \lambda_2 = \lambda$) will be sufficient and the grid search can be one-dimensional.

3.2 Toward higher dimensions

Although this scheme will work, it can be wasteful with memory and computation time, especially for higher dimensions. For example, in three dimensions (with common $n$ and $c$), the basis would have $n^3$ rows and $c^3$ columns and most practical applications would not fit in the memory of the average PC. To avoid this bottleneck, we use the grid algorithms of Eilers et al. (2006). An example is the computation of the mean $\mu = \text{vec}(\mathbf{M})$ in two dimensions. Instead of computing it as a vector, as in (9), we can use the following equivalence

$$\mu = E(y) = B_2 \otimes B_1 \theta = B \theta,$$

$$\mathbf{M} = E(Y) = B_1 \Theta B_2'. \quad (13)$$

This avoids the construction of the large Kronecker product basis, saving space and time. This scheme can be extended to 3 dimensions or more, by recursive re-arrangement of multi-dimensional arrays as matrices and pre-multiplication by the one-dimensional bases. In a similar way the multi-dimensional “inner products” $B'WB$ can be obtained efficiently. The key observation is that in a 4-dimensional representation the elements of $B'WB$ can be written

$$f_{jkk'k'} = \sum_h \sum_i w_{hi} b_{ij} b_{hk} b_{ij'} b_{hk'}, \quad (14)$$

which can be rearranged as

$$f_{jkk'k'} = \sum_h b_{hk} b_{hk'} \sum_i w_{hi} b_{ij} b_{ij'}. \quad (15)$$

By clever use of row-wise tensor products and switching between four- and two-dimensional representations of matrices and arrays, one can completely avoid the construction of the large tensor-product basis $B$. In addition to saving memory, computations are sped up by orders of magnitude. There is no room for details here, see Eilers et al. (2006) and Currie et al. (2006).
4 Illustrative examples

Figure 1 shows two-dimensional density contours for the well-known Old Faithful data. The number of Poisson observations is on a $100 \times 100$ histogram.
togram grid, and a $13 \times 13$ grid of tensor products of cubic B-splines were used to construct the two-dimensional density. The order of the difference penalty is 3. We see that AIC is minimized with relatively light smoothing ($\lambda_1 \approx 0.003$) along the rows (waiting times) and with heavier smoothing ($\lambda_2 \approx 1000$) along the columns (duration). Figure 2 shows the result of isotropic smoothing, yielding a compromise in the single $\lambda \approx 0.02$. The large difference between the two $\lambda$s for anisotropic smoothing might look surprising. The reason is that the conditional distributions of waiting time given (the previous) duration is always unimodal, not much different from a normal distribution. This allows a large weight for a third order penalty without destroying a proper fit to the data. On the other hand the conditional distribution of the (previous) duration given the waiting time is bimodal or rather skewed on part of the data domain. The weight of “horizontal” penalty cannot therefore be large.

As a second example, Figure 3 displays as scatterplot of gene expressions (on a log$_{10}$ scale) as measured by two microarrays. The data pairs are plotted using a simple “dot” symbol. Only a random selection of 1000 observa-

FIGURE 3. Presentation of microarray data scatterplot as a density. Upper: scatterplot of selection of raw data (left); optimal isotropic penalty parameter using AIC (right). Lower: image of optimal density (left); square root density (for improved visualization) (right).
tions is shown, to prevent a large part of the cloud of pints to become completely black. As discussed by Eilers and Goeman (2004), scatterplots with many dots are difficult to judge, because symbols at the boundaries attract too much attention from the observer, making the “cloud of points” look wider than it really is. Eilers and Goeman presented a very fast but simple smoother, with a goal of an attractive display of dense scatterplots. The density smoother proposed here can be used as an improvement, especially when $\lambda$ is optimized by AIC. The results presented were also obtained with a $100 \times 100$ histogram grid and $13 \times 13$ tensor products of cubic B-splines. The optimal penalty weight value for isotropic smoothing is $\lambda \approx 0.0002$.

The result of the smoother is presented as a gray-scale image of the height of the density. Depending on the medium used (computer screen, projector or printer) the effective dynamic range can be quite small. Therefore we also present a gray scale plot of the square root of the density.

5 Discussion

We have shown that tensor products of P-splines and Poisson regression of histogram data leads to an effective density estimator. As discussed in Eilers and Marx (1996), the position and number of bins (i.e. the location and width of the histogram grid) makes little difference to the final fit, provided that the grid is chosen to be sufficiently fine. There is no room here to illustrate that the same holds for two-dimensional smoothing. It is remarkable that a very sparse two-dimensional histogram leads to a very good density estimate. The Old Faithful data set has 272 observations. The average count per 2-D histogram bin is thus less than 0.03! As a practical guideline in two dimensions, we recommend that users (who are not concerned with computation time) start with $100 \times 100$ histogram bins, $20 \times 20$ B-splines, and a second or third order penalty.

Some nice features of one-dimensional P-spline density estimation also carry over to higher dimensions. For one, there is a conservation of moments, for any $\lambda$, within each row, column, layer and so forth (depending on the dimension of the histogram). For example within any row, a penalty of order $d = 1$ yields equality of the sum of the observed and the sum of fitted counts. When $d = 2$, e.g. within any row, the previous condition holds and the mean of the observed data equals that of the fitted counts. Further, when $d = 3$, the previous two conditions hold with an additional preservation of variance between observed and fitted data.

The P-spline density estimator is also not affected by unpleasant boundary effects, so familiar to kernel smoothers. In fact sharp specialized boundaries are encouraged with the P-spline approach. Especially for isotropic smoothing, optimization of AIC, using a simple grid search, is efficient. Computation, for one value of $\lambda$ (in Matlab), takes less than a second on a 1000 MHz PIII computer. The time to compute the histogram is negligible, even for many observations.
Minimization of AIC is not the only possible approach to optimal smoothing. One can interpret models with penalties as mixed models and attempt to estimate the variance of the mixing distribution (Ruppert, Wand and Carroll, 2003). Durbán et al. (2006) investigated this approach to multidimensional density estimation. Alternatively, a purely Bayesian approach is also possible. The penalty is seen as the logarithm of a prior density of differences of the coefficients of the B-spline tensor products. Efficient simulation, using an optimized Langevin-Hastings algorithm, is possible. See Lambert and Eilers (2006), extending their 2005 work.

References


Approximations for the MSE of random effects in a GLMM estimated by PQL

Escaramís G.\(^1\) and Carrasco J.L.\(^1\)

\(^1\) Dept. Salut Pública, Universitat de Barcelona, Spain

Keywords: Best linear unbiased predictor; mean squared error; generalized linear mixed model; penalized quasi-likelihood.

1 Introduction

Spatial disease mapping concerns the analysis of the geographical distribution of a particular disease. The data to be analyzed often consist of counts of diseased people within sub-regions of an area under study, such as health sections or administrative areas. In this kind of study it is usually of interest to identify which sub-regions show extreme risks. When studying rare diseases or small areas the use of direct risk measures can yield large standard errors and hence a low power to detect extreme risks. Thus, to overcome this issue, smoothing methods such as the generalized linear mixed model (GLMM) can be applied. The use of GLMM can accommodate the overdispersion arising from unmeasured risk factors (risk heterogeneity) or when sub-regions share environmental factors (spatial correlation) by incorporating sub-regions as random effects with a prior distribution depending on variance components. Let us assume we have a response \(Y_j\) of disease counts, with \(E_j\) being the number of expected counts based on standard population rates and \(\gamma_j\) the underlying disease relative risk for the \(j\)th sub-region, \(j = 1, \ldots, N\). Now, consider the conditional mean response given a random effect vector, \(\mu_j^b = E(Y_j|b)\). We can thus specify a GLMM using a log-link function; 
\[
\log(\mu_j^b) = \log(E_j) + \alpha + b_j,
\]
where \(\alpha\) is the baseline log-rate across all sub-regions and \(b_j = \log(\gamma_j)\) is the random effect for the \(j\)th sub-region. Hence, the response variables conditional on the random effects are assumed to be independent with Poisson distribution \(Y_j|b_j \sim \text{Poisss}(\mu_j^b)\). Furthermore, to take into account the spatial correlation of the data and the heterogeneity of the risk, the following conditional autoregressive (CAR) prior distribution (Besag, York and Mollié, 1991) can be specified; \(b = (b_1, \ldots, b_N) \sim MVN(0, \Sigma(\theta))\), \(\Sigma(\theta) = \Sigma_0 + \Sigma_1\), where \(\Sigma_0 = I + Q^{-1}\), and \(\Sigma_1\) and \(\Sigma_2\) are the variance components reflecting, respectively, the variability that is not structured and the variability which is spatially structured, and where
\( \mathbf{I} \) is the identity matrix with dimension \( N \) and \( \mathbf{Q} \) is a neighbourhood matrix which defines the spatial structure. A common choice for \( \mathbf{Q} \) is defining the neighbourhood by sub-region adjacency, in which case the diagonal elements of the matrix are equal to the number of neighbours of the corresponding sub-region and the off-diagonal elements are equal to \(-1\) if the corresponding sub-regions are neighbours and \(0\) otherwise. Inference about random effects reflects whether there are some sub-regions showing extreme risks or not. The GLMMs can be fitted via full maximum analysis based on the joint marginal distribution, which requires numerical integration techniques. A common approach is the fully Bayesian (FB), which obtains estimates via Markov Chain Monte Carlo (Bernardinelly and Montomoly, 1992). This approach is able to take into account the uncertainty associated with the variance component estimates specifying vague hyperpriors. Alternatively, penalized quasi-likelihood (PQL) has been proposed as a simpler method which approximates maximum likelihood (Breslow and Clayton, 1993). PQL yields adequate point predictions for disease risks based on two-stage predictors, by first obtaining the best linear unbiased predictor (BLUP), assuming that the variance components are known, and then replacing the variance components in the predictor with their estimates. Thus, the uncertainty that arises from having to estimate variance components is not taken into account and the Mean Squared Error (MSE) of these predictors might be underestimated. The aim of this study is to evaluate two analytical approaches for the MSE of the random effects predictions when the model is estimated via PQL. The two approaches incorporate the fact that variance components are estimates rather than parameters. We apply these approaches to the particular GLMM described in a case example study.

2 Ways of estimating the MSE of the Random Effects when using PQL

A direct approximation to the PQL MSE (MSE0) estimate is that which does not take into account the uncertainty associated with the fact that we replace the variance component parameters by their estimates

\[
\text{MSE0} \left[ \hat{b}_j(\hat{\theta}) \right] = c^T \{ \Sigma(\hat{\theta}) - \Sigma(\hat{\theta}) \mathbf{Z}' \mathbf{V}^{-1}(\hat{\theta}) \mathbf{Z} \Sigma(\hat{\theta}) + \\
+ \Sigma(\hat{\theta}) \mathbf{Z}' \mathbf{V}^{-1}(\hat{\theta}) \mathbf{X}_{var}(\hat{\alpha}) \mathbf{X}' \mathbf{V}^{-1}(\hat{\theta}) \mathbf{Z} \Sigma(\hat{\theta}) \}
\]

where \( \mathbf{V}(\hat{\theta}) \) is the data variance-covariance matrix, \( \mathbf{Z} \) is the random effects design matrix and \( c \) is a vector with components \( c_k = 1 \) if \( k = j \) and \( c_k = 0 \) otherwise. The first proposal for the MSE (MSE1) to reflect the uncertainty that arises from having to estimate variance components instead of plug-in true values is derived from the study of Kackar and Harville (1984), where the authors perform a general approximation
to the MSE of a two-stage predictor in a general linear mixed model;
\[ \text{MSE}_1 \left[ \hat{b}_j(\hat{\theta}) \right] \approx \text{MSE}_0 \left[ \hat{b}_j(\hat{\theta}) \right] + \text{tr} \left[ \text{var} \left( \frac{\partial \hat{b}_j}{\partial \theta} \right) \text{B}(\theta) \right], \]
where \( B(\theta) \) is either the Fisher informative matrix of \( \hat{\theta} \) or some approximation to its MSE. The second approach (MSE2) is based on the results obtained by Prasad and Rao (1990), where they suggest a further approximation and apply it to three special cases of small-area linear mixed models;
\[ \text{MSE}_2 \left[ \hat{b}_j(\hat{\theta}) \right] \approx \text{MSE}_0 \left[ \hat{b}_j(\hat{\theta}) \right] + \text{tr} \left[ (\partial \hat{b}_j^* / \partial \theta)(V(\theta)(\partial \hat{b}_j^* / \partial \theta) \text{B}(\theta)) \right], \]
where \( b_j^* = \Sigma(\theta)Z^TV^{-1}(\theta) \).

3 Type I diabetes Mellitus Incidence Data

To illustrate the different approaches presented in the previous section for detecting extreme risks we analysed type I diabetes mellitus incidence data from Catalonia for the period 1989-1998. The data consist of the incidence of type I diabetes mellitus in people under 30 and the expected counts for the 41 administrative local areas in Catalonia. In order to explore whether the overdispersion observed when we fitted an independent Poisson model could in part be spatially structured in terms of sub-region adjacency, we calculated the adjusted rank adjacency statistic for spatial clustering; this indicated a possible degree of dependency among adjacent sub-regions. We then fitted the GLMM via PQL to estimate the MSE and also via FB specifying the following vague hyperpriors for the inverse of the variance components: \( \tau_H \sim \text{Gamma}(0.5, 0.0005) \) and \( \tau_S \sim \text{Gamma}(0.5, 0.0005) \). Figure 1 compares relative risk predictions for the two approaches as well as for the three approximations to the MSE of the relative risks via PQL vs FB. Relative risk predictions using FB and PQL approaches are nearly identical, but the PQL MSE0 is slightly shrunk compared to the FB, which may mislead us to detect extreme risks. In contrast, accommodation of both the approximations (MSE1 and MSE2) produce larger MSE, and at the same time both turn out to be slightly larger than the FB counterpart; thus inferences based on these MSE approaches will be more conservative.

4 Simulation study

To investigate whether MSE1 and MSE2 reflect the uncertainty associated with the relative risk predictions when fitting the specified GLMM we carried out a simulation study. We first set \( \sigma_H^2 \) and \( \sigma_S^2 \) equal to the GLMM point estimates to obtain 1000 realizations of the random effects, \( b_j \), from the appropriate multivariate normal distribution. We then calculated 1000 samples of diabetes incidences in each administrative area, \( Y_j \), from independent Poisson distributions setting \( \alpha \) equal to its point estimate, \( E_j \), and \( b_j \). Furthermore, we explored the interval prediction coverage of the form
exp\{\hat{b}_i^j \pm z_{0.95} \sqrt{MSE_i^j}\}. We found that MSE1 and MSE2 produced interval predictions vaguely exceeding the 90% nominal coverage, while MSE0 produces the opposite effect, corroborating the hypothesis of being more conservative when taking into account such uncertainty.

Summing up, eventhough in our example the MSE0 provides desirable coverages rates, the MSE1 and MSE2 slightly improve them. A rationale of this result is that the variance components in our data have a relatively low standard error, thus a low impact in the random effect MSE estimates. However, these differences can increase in other data with a greater uncertainty in the variance component estimation.

References


Modelling non-Gaussian time series with a mixture copula transition model

Gabriel Escarela¹, Ramsés H. Mena² and Alberto Castillo-Morales¹

¹ Address of first Department de Matemáticas, Universidad Autónoma Metropolitana, Unidad Iztapalapa, AT-321, Av. San Rafael Atlixco No. 186, Col. Vicentina, C.P. 09340, México D.F., Mexico, ge@xanum.uam.mx
² Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM, A.P. 20-726. 01000 México D.F., Mexico.

Abstract: We present a general parametric class of transitional models of order p. Specifically, we extend the work of Raftery (1985), giving rise to a more generalised autoregressive mixture model that allows for covariates. In these models the conditional distribution of the current observation given the present and past history is a mixture of conditional distributions, each of them corresponding to the current observation given each one of the p-lagged observations. Such conditional distributions are constructed using bivariate copula models which allow for a rich range of dependence suitable to model non-Gaussian time series. This likelihood-based methodology provide with a convenient way to study time series data which may not be Gaussian in a general framework.

Keywords: Autoregressive models; Markov regression models; mixture transition distribution model; non-Gaussian data; Poisson regression model.

1 Introduction

There has been a growing interest in longitudinal studies in the biomedical field. In such studies, it is preponderant to consider time series structures which are capable to capture both the marginal behaviour and the dependence between the responses. When it comes to linear Gaussian time series, there exists a very rich and well established literature concerning such structures. In many applications, however, the Gaussian assumption does not hold implying the need of more general models. Transition models provide with a convenient alternative to study longitudinal data in a general framework. In this study, we shall be interested in the construction of transition models where both past outcomes and past and present covariates are considered in the model.

To stablish context, consider first-order discrete time stationary Markov processes. If the state space is continuous, the process can be constructed from a given bivariate distribution \( F(y_1, y_2) = \Pr\{Y_1 \leq y_1, Y_2 \leq y_2\} \), which
corresponds to a jointly continuous random vector \((Y_1, Y_2)\) with the two univariate marginal distributions both equal to the stationary distribution. The transition distribution, defined as \(F_{2|1}(y_2 \mid y_1) = \Pr\{Y_2 \leq y_2 \mid Y_1 = y_1\}\), can be computed as

\[
F_{2|1}(y_2 \mid y_1) = \frac{\partial F(y_1, y_2)}{\partial y_1} / \frac{\partial F(y_1, \infty)}{\partial y_1},
\]

where \(F(y_1, \infty)\) denotes the marginal distribution function of \(Y_1\). When the state space of the process is a finite or countable set, the transition distribution is given by

\[
F_{2|1}(y_2 \mid y_1) = \sum_{z \leq y_2} \frac{f(y_1, z)}{f_1(y_1)}.
\]

where \(f(y_1, y_2)\) is a joint mass function with both marginal probability functions equal to \(f_1(\cdot)\).

For the model we will propose in the next section, we will require the construction of bivariate distributions whose marginals are equal. Although, there are many bivariate distributions in the literature, few of these share the property of having the same marginal distributions.

1.1 Copulas

Copula models are classes of bivariate distribution functions specified in terms of the marginal distribution functions and a copula function, which is a continuous bivariate distribution function on \((0, 1)^2\) with uniform marginals. An attractive characteristic of the copula class is that the elimination of the marginals through the copula helps to model and understand dependence structures effectively, as the dependence has no relationship with the marginal behaviour of individual characteristics. Thus, the functional dependence can be studied without the need of specifying the marginal distributions.

In the continuous context, it is possible to construct a bivariate distribution function in terms of two marginal distribution functions and a copula that allows for dependence relationships among the individual random variables. Thus, the bivariate distribution copula \(C\) of \((Y_1, Y_2)\) is found by making marginal probability integral transforms on \(Y_1\) and \(Y_2\) so that the joint distribution function is expressed as follows:

\[
F(y_1, y_2) = C[F_1(y_1), F_2(y_2)].
\]

Provided that \(F_1(y_1), F_2(y_2)\) and the copula \(C(v_1, v_2)\) are differentiable, the joint density of \((Y_1, Y_2)\) can be expressed as

\[
f(y_1, y_2) = f_1(y_1)f_2(y_2) \times c[F_1(y_1), F_2(y_2)],
\]
where \( f_1(y_1) \) and \( f_2(y_2) \) are the marginal density functions, and

\[
c(v_1, v_2) = \frac{\partial^2 C(v_1, v_2)}{\partial v_1 \partial v_2}, \quad (v_1, v_2)^T \in (0, 1)^2,
\]

is the corresponding copula density function. It follows that the conditional density function of \( Y_2 \) given \( Y_1 \) can be conveniently written as:

\[
f_{2|1}(y_2 \mid y_1) = f_2(y_2) \times c[F_1(y_1), F_2(y_2)]. \tag{1}
\]

### 1.2 Two Families of Copulas

Choosing a copula that allows for a wide range of dependence is an important issue. Next, we present two important families of bivariate copula models which have a number of desirable properties.

**Gaussian Family.** The bivariate distribution function belonging to the Gaussian copula has the form

\[
C_\rho(v_1, v_2) = \Phi_2[\Phi^{-1}(v_1), \Phi^{-1}(v_2)], \quad (v_1, v_2)^T \in (0, 1)^2,
\]

where \( \Phi_2(\cdot, \cdot) \) is the cdf of a bivariate Gaussian distribution with mean \((0, 0)^T\) and covariance matrix \( R \) equal to an \( 2 \times 2 \) non-singular matrix with the off-diagonal elements equal to \( \rho \), with \( \rho \in (-1, 1) \), and the diagonal elements equal to one, and \( \Phi^{-1}(\cdot) \) is the inverse function of the standard Gaussian cumulative distribution. The use of the bivariate Gaussian copula is appealing since it encodes dependence in the same way that the bivariate normal distribution does using the dependence parameter \( \rho \), with the difference that it does so for random variables with any arbitrary marginals.

**Positive Stable Family.** The bivariate distribution function presented by Hougaard takes the form

\[
C_\alpha(v_1, v_2) = \exp \left\{ - \left[ (-\log v_1)^{1/\alpha} + (-\log v_2)^{1/\alpha} \right]^\alpha \right\}, \quad \alpha \in (0, 1). \tag{2}
\]

This copula is useful for modelling positive dependence. Small values of \( \alpha \) provide large positive dependence between \( Y_1 \) and \( Y_2 \), whilst large values of \( \alpha \) provide associations close to independence.

### 2 The transition model

#### 2.1 The First-order transition model

Using the conditional copula density function given by equation (1) in terms of a given copula density function and a given margin, \( Y_t \sim f \), one can construct a transitional model for continuous responses straightforwardly. If, for instance, one wishes to obtain the first-order Gaussian autorregressive
model in terms of the transition copula model, one simply needs to provide the Gaussian copula density function \( c_p(\cdot, \cdot) \) (Song, 2000), the standardised margin distribution function \( F(y_t) = \Phi((y_t - \beta^T x_t)/\sigma) \) and the marginal density \( f(y_t) = (2\pi\sigma^2)^{-1/2} \times \exp\{- (y_t - \beta^T x_t)^2/(2\sigma^2)\} \).

When first-order transitional models are for discrete responses, a conditional probability function can also be obtained to define the transition distribution of \( Y_t \mid Y_{t-1} \) when the distribution of \( Y_t \) is given. If \( f(y_t) = \Pr\{Y_t = y_t\} \) represents the marginal probability function of \( Y_t \), the family of transition distributions of \( \{Y_t\} \) can be characterised by using the discretised bivariate copula and the discrete distribution function \( F(y_t) = \sum_{z \leq y_t} f(z) \) as follows (Joe, 1996):

\[
F_{21}(y_t \mid y_{t-1}) = \{C[F(y_{t-1}), F(y_t)] - C[F(y_{t-1} - 1), F(y_t)]\} / f(y_{t-1}).
\]

Using the conditional copula density function given by equation (1) in terms of a given copula density function and a given margin, \( Y_t \sim f \), one can construct a transitional model for continuous responses. If, for instance, one wishes to obtain a first-order Gaussian autorregressive model with \( Y_t \sim N(\beta^T x_t, \sigma^2) \), where \( x_t \) is a vector of explanatory variables observed at time \( t \) and \( \beta \) is the corresponding vector of unknown regression coefficients, one simply needs to provide the Gaussian copula density function \( c_p \) (Song, 2000), the standardised margin distribution function \( F(y_t) = \Phi((y_t - \beta^T x_t)/\sigma) \) and the marginal density \( f(y_t) = (2\pi\sigma^2)^{-1/2} \times \exp\{- (y_t - \beta^T x_t)^2/(2\sigma^2)\} \).

When first-order transitional models are for discrete responses, a conditional probability function can also be obtained to define the transition distribution of \( Y_t \mid Y_{t-1} \) when the distribution of \( Y_t \) is given. If \( F(y_t) \) represents the marginal cdf of \( Y_t \), the family of transition distributions of \( \{Y_t\} \) can be characterised by the discrete transition density function \( f_{21}(y_t \mid y_{t-1}) = \Pr\{Y_t = y_t \mid Y_{t-1} = y_{t-1}\} \) as follows (Joe, 1996):

\[
f_{21}(y_t \mid y_{t-1}) = \{C[F(y_t), F(y_{t-1})] - C[F(y_t), F(y_{t-1} - 1)] - C[F(y_t - 1), F(y_{t-1})] + C[F(y_t - 1), F(y_{t-1} - 1)]\} / f(y_{t-1}).
\]

### 2.2 The Mixture transition distribution model

In principle, it is possible to extend the transition copula models described above to higher-order representations. In practice, however, this can be quite cumbersome and computationally expensive. A more parsimonious approach than a fully parametric Markov chain can be the mixture transition distribution (MTD) model introduced by Raftery (1985) which is characterised by the following conditional density

\[
f(y_t \mid y_{[t,p]}) = \sum_{k=1}^{p} \omega_k f_k(y_t \mid y_{t-k}). \tag{3}
\]
where $y^{[t;p]} = (y_{t-1}, ..., y_{t-p})$, $\sum_{k=1}^{p} \omega_k = 1$, $\omega_k \geq 0$, $k = 1, ..., p$, and $f_k(y_t \given y_{t-k})$ denotes the one-step ahead transition density corresponding to lag $y_{t-k}$. Raftery showed that the lagged bivariate distributions satisfy a system of matrix equations similar to the Yule-Walker equations, and that the past values $Y_{t-1}, ..., Y_{t-p}$ do not interact among them in their effect on the conditional distribution of $Y_t$ given the past. Raftery’s MTD model does not limit to a specific parametric form for the one-step conditional density functions $f_k(y_t \given y_{t-k})$. In general, it is not straightforward to incorporate a given margin and covariate information. To construct each of the one-step conditional densities in equation (3), we employ the copula transition models described above by choosing both a family of copulas and a family of distributions for the underlying marginals. In order to both choose a suitable parametric family for the marginal distributions and allow for the effect of covariates, it may be convenient to define the marginals in terms of the generalised linear model.

3 Inference

Assuming that the observation times of the responses $\{y_t, t = 1, ..., n\}$ are equally spaced, the likelihood of the transition model of order $p$ can be expressed as

$$L = f(y_1; x_1) \prod_{k=2}^{n} f_k(y_k \given H^{[k,j]}),$$

where $j = \min(k-1, p)$, and $H^{[k,j]} = \{y_{k-1}, ..., y_{k-j}; x_k, x_{k-1}, ..., x_{k-j}\}$. The mixture transition model is not linear, so numerical techniques need to be used in order to find the maximum likelihood estimators. Our approach to applying the transition models is to determine the order of the Markov chain prior to inference about the regression and dependence parameters. In this study we consider the Bayesian information criterion (BIC) as our main model choice criterion. Once the order of the model is fixed, the process of finding a parsimonious model can follow standard ideas based on the likelihood ratio; for example, that used in the theory of generalised linear models.

The assumptions made about a particular transitional model may be checked by calculating the conditional randomised quantile residuals proposed by Dunn and Smyth (1996). Under the assumed model, such residuals are exactly normal and are found by inverting the fitted conditional distribution function for each response values and finding the equivalent standard normal quantile. Thus, some simple plots for checking that the randomised quantile residuals are observed values of independent and normally standard distributed random variates should indicate the quality of the fit.
TABLE 1. Deviances of various transitional mixture models containing trend and both seasonal effects using the positive stable copula and Poisson margins.

<table>
<thead>
<tr>
<th>Order</th>
<th>$-2 \times \text{loglik}$</th>
<th>$n_p$</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independence</td>
<td>543.55</td>
<td>6</td>
<td>555.55</td>
<td>574.29</td>
</tr>
<tr>
<td>1</td>
<td>532.20</td>
<td>7</td>
<td>546.20</td>
<td>568.06</td>
</tr>
<tr>
<td>2</td>
<td>488.17</td>
<td>15</td>
<td>518.17</td>
<td>565.02</td>
</tr>
<tr>
<td>3</td>
<td>486.44</td>
<td>23</td>
<td>532.44</td>
<td>604.29</td>
</tr>
<tr>
<td>4</td>
<td>485.11</td>
<td>31</td>
<td>547.11</td>
<td>643.96</td>
</tr>
</tbody>
</table>

$n_p$ = number of parameters

4 Illustration

To illustrate the methods, we analyse the time series of monthly number of cases of poliomyelitis reported by the U.S. Centers for Disease Control for the years 1970 to 1983 published by Zeger (1988). The goal of the analysis is to assess whether the incidence of polio has been decreasing since 1970 in presence of any seasonality. In this study the same set of covariates used by Zeger (1988) is chosen, taking into account the intercept, $x_0 = 1$, trend, $x_1 = t/1000$, seasonal effects in the form of annual frequencies, $x_2 = \cos(2\pi t/12)$ and $x_3 = \sin(2\pi t/12)$, and of semi-annual frequencies, $x_4 = \cos(2\pi t/6)$ and $x_5 = \sin(2\pi t/6)$.

We fit the discrete transition regression model employing the positive stable copula and Poisson marginal distributions. Specifically, each of the $k$-th one-step conditional distributions in the mixture model are constructed by defining $C_k(v_1, v_2) \equiv C_{\alpha_k}(v_1, v_2)$, where $C_{\alpha_k}$ is the positive stable copula with dependence parameter $\alpha_k$ as defined in equation (2), and $F_k(y_t) = \sum_{z \leq y_t} f_k(z)$, where $f_k(y_t) \equiv \text{Poi}(y_t; \mu_k(\eta_t))$ is the Poisson probability function with mean $\mu_k(\eta_t)$ defined as

$$f_k(y_t) = \exp\{-\mu_k(\eta_t)\} \frac{\mu_k(\eta_t)^{y_t}}{y_t!} \times I_{\{0, 1, \ldots\}}(y_t). \quad (4)$$

Here, we assume a log-linear model for each marginal expectation, i.e. $\log \mu_k(\eta_t) = \beta_k^T x_t$.

In order to find the adequate order of the MTD model, we fit the transition mixture model to the Polio data in presence of all covariates for various values of $p$. Table 1 shows the resulting deviances and the corresponding Bayes factors. It is possible to observe a dramatic decrease in deviance when comparing the mixture models with a simpler model such as the log-linear Poisson regression model which assumes independent responses. Under the criteria mentioned above, the model based on two mixtures fits better the data than any other model presented here. Thus, we choose the second-order model.
We observed that the estimator of the dependence parameter of the one-step transition probability component in the second-order MTD model, \( f(y_t \mid y_{t-2}) \), was very close to 1 (\( \hat{\alpha}_2 = 0.971 \)), suggesting that the independent copula could be more suitable for this component. When it came to fitting the data using independence for the two-step transition probability function, i.e. when \( f(y_t \mid y_{t-2}) = f_2(y_t) \), the change in deviance was negligible. The resulting model can then be interpreted as a mixture of the current and first lag histories, which is not strictly a second order transition model.

Consequently, we proceed to fit the transitional model with the following specification:

\[
\begin{align*}
f(y_t \mid H^{t,[2]}) &= \omega f_1(y_t; z_t) + (1 - \omega) f_{2|1}(y_t \mid H^{t,[1]}),
\end{align*}
\]

where \( 0 \leq \omega \leq 1 \),

\[
\begin{align*}
f_{2|1}(y_t \mid H^{[t,1]}) &= \{ \alpha C_\alpha [F(y_t; x_t), F(y_{t-1}; x_{t-1})] - \\
&\quad C_\alpha [F(y_t; x_t), F(y_{t-1} - 1; x_{t-1})] - \\
&\quad C_\alpha [F(y_t - 1; x_t), F(y_{t-1}; x_{t-1})] + \\
&\quad C_\alpha [F(y_t - 1; x_t), F(y_{t-1} - 1; x_{t-1})] \}/f(y_{t-1}; x_{t-1}),
\end{align*}
\]

\( f_1(y_t; z_t) \) and \( f(y_t; x_t) \) are probability functions corresponding to Poisson distributions with mean \( \exp\{\beta_1^T z_t\} \) and \( \exp\{\beta_2^T x_t\} \) respectively, and \( F(y; x_t) = \sum_{h \leq y} f(h; x_t) \). Notice that \( z_t \) may contain some or all of the covariates in \( x_t \), as well as other covariates not included in \( x_t \); we want to find and fit satisfactory models for both components.

We found that for \( f_1 \) trend and annual effects are significant. On the other hand, only semi-annual effects are significant for \( f_{2|1} \). Table 2 shows the MLE’s and standard errors of the most parsimonious model obtained after backward elimination, which we will refer to as the best-fitting transition model. Notice that the estimates of the dependence parameter \( \alpha \) and the mixture parameter \( \omega \) are given in terms of the logit function, that is \( \text{logit}(x) = \log[x/(1 - x)] \). Table 2 also displays the corresponding observed Bayes factors which, compared to those reported in Table 1, show a reduction.

The negative coefficient corresponding to trend in \( f_1(y_t; z_t) \) and its statistical significance (p-value=0.0005) indicate a decreasing trend in the conditional mean after accounting for both the seasonality and the dependence structure exhibited by the data, suggesting that the average monthly number of cases of poliomyelitis decreases over time.

After carrying out a residual analysis, which consists of kernel density, normal Q-Q plot, autocorrelation and partial autocorrelation functions of the randomised quantile residuals, we found that the proposed model appears to fit the data reasonably well, though there is a small deviation from the theoretical line in the upper tail of the distribution. There is no indication of correlation between the residuals.
TABLE 2. Estimates of the coefficients and standard errors for the best-fitting transition model characterised by equation (5).

| Parameter            | $I_1(y_t; z_t)$ | $I_2(1 | y_t H^{[y,t]})$ |
|----------------------|-----------------|------------------------|
| Intercept            | 0.646           | -1.003                 |
| $t/1000$             | -7.022          | -                      |
| cos($2\pi t/12$)     | 0.113           | -                      |
| sin($2\pi t/12$)     | -0.531          | -                      |
| cos($2\pi t/6$)      | -                | 2.398                  |
| sin($2\pi t/6$)      | -                | -0.721                 |
| logit $\alpha$       | -                | 1.755                  |

<table>
<thead>
<tr>
<th>Mixture Parameter</th>
<th>Estimate</th>
<th>S.E.</th>
<th>Deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>logit $\omega$</td>
<td>1.210</td>
<td>0.373</td>
<td>$-2 \times \log1k = 496.31$</td>
</tr>
</tbody>
</table>

The transition mixture model presented here is well suited for the parametric representation of high-order dependencies between successive observations of a random variable. Since there can be considerable differences in the shapes of conditional distributions generated by the copula model, even when the fitted marginal distribution and estimated correlations are similar, it is important to choose correctly a family of copulas and a marginal distribution for each conditional distribution in the mixture model.

References


Markov modelling of temporally-ordered functional response data

John S. Fenlon$^1$ and Malcolm J. Faddy$^2$

$^1$ Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK
$^2$ School of Mathematical Sciences, QUT, Brisbane, Qld 4001, Australia

1 Introduction

Functional response is important in understanding the dynamics of predator-prey and parasitoid-host systems - it is essentially the interpretation of a bio-assay system in which individual predators (or parasitoids) have access to fixed numbers of prey for a given period of time. The classical approach to the problem has entailed the use of mechanistic models to interpret the data, but more recently several papers have argued that the use of simple logistic regression is both more consistent with the nature of the data, and allows for the stochastic variation inherent in the system. Nevertheless, both the classical approach and this newer interpretation focus only on the modelling of means, and ignore the variability of the data. A further difficulty, usually overlooked, is that many published data sets display over-dispersion which itself may be a function of prey density.

Fenlon and Faddy (2006) show how the beta-binomial model can be readily adapted to such data, but also present some new models based on stochastic counting processes. The new method models the instantaneous rate, $\lambda_n$, at which new prey are eaten given that $n$ prey have already been eaten as a product of two components $a_n$ and $b_{N-n}$: the first could be a simple increasing function of $n$ reflecting the predators’ appetite, while the second will be a function of the available prey $N-n$ (where $N$ is the initial number of prey available). The corresponding equations are solved numerically using MATLAB software. It is argued that the counting process models offer richer insights into the predation process than do other more ‘descriptive’ models such as those based on the beta-binomial. The findings highlight an interesting apparent paradox: that inferences can be made about changing predation rates without an explicit time response. That, however, is implicitly assumed in the class of models employed, in which assumptions are made about a process of which only the endpoint is observed.

Here we extend the model and make the $\lambda_n$’s explicitly time-dependent, $\lambda_n(t)$, and estimate them as piece-wise constant temporal rates. The model is tested on data, presented in Toussidou et al. (2006), where the functional
response of a parasitoid-host system was examined in an experiment in which independent assays were measured at $\frac{1}{2}$ h, 1, 2, 5 and 24 h. It is shown here how the rates of parasitisation vary over time, and how this temporal model provides a better description of progress to the 24 h endpoint than does simple modelling of the 24 h data alone. There is an obvious distinction between predator-prey and parasitoid-host models in that the predator consumes its prey whereas the parasitoid does not, making hosts available for super-parasitism. The same type of assay system exists for both functions, though there are differences in interpretation in classical ecological models. A good reference is Juliano (2001).

2 Data

The data are from a series of experiments in which colonies of aphids were challenged by individual parasitoids. Four replicates of a $5 \times 5$ factorial, five nominal density levels with actual aphid numbers ranging from $< 10$ to 500 aphids, and five foraging times: $\frac{1}{2}$, 1, 2, 5 and 24 h. Each assay took place in an individual cage containing a whole single-stemmed chrysanthemum plant. The total number of aphid mummies collected from each treatment was used to provide a measure of parasitism at different combinations of foraging time and aphid densities. A visual representation of the data is given in Figure 1 and the following basic observations can be made:

1. In the $\frac{1}{2}$ hour assay, two parasitoids laid no eggs, and several more laid very few, but in all other assays mummies were recovered.

2. For foraging times up to and including 2 h time appears to be limiting, i.e. the response becomes quite flat above a certain availability; whereas, at the two longest foraging times (5 and 24 h) the response appears more linear, showing a strong association between predation and availability.

3. Some of the variability apparent at 5 h has disappeared by 24 h.

3 Modelling

Faddy and Fenlon (1999) used stochastic counting processes $\{X(t); t \geq 0\}$, introduced in Faddy (1997), to describe the number of nematodes invading a host. Here we let $X(t)$ represent the number of hosts parasitised at time $t$, and $\lambda_n$ the instantaneous rate at which new hosts are parasitised given that $n$ hosts have already been attacked. Then, transition probabilities:

\[ P\{X(t + \delta t) = n + 1 \mid X(t) = n\} = \lambda_n \delta t + o(\delta t) \]
with $X(0) = 0$ and $\lambda_N = 0$ model the process of parasitisation. The value of $X(t)$ after a fixed period of time, which can be taken without loss of generality to be 1, is considered as the response, i.e. the number of hosts parasitised. If $\lambda_n = a(N - n)$, the distribution of $X(1)$ is binomial with $p = 1 - e^{-a}$. For generalisations with $\lambda_n = a_n(N - n)$, distributions of $X(1)$ with over-dispersion arise from increasing sequences $a_n, n = 0, 1, 2, \ldots, N$. With this in mind, Faddy and Fenlon (1999) considered forms in which $a_n$ increased exponentially or logistically. Fenlon and Faddy (2006) introduced modified forms for $\lambda_n$ which allow for non-exchangeability of the prey by replacing the term $(N - n)$ by a more general function of $(N - n)$.

In fitting these models it is necessary to calculate the probabilities $p_n(t) = P\{X(t) = n\}$ for $n = 0, 1, \ldots, N$. The Kolmogorov forward differential equations for these probabilities are given by:

$$\frac{dp_0(t)}{dt} = -\lambda_0 p_0(t), \text{ with } p_0(0) = 1,$$
and for \( n = 1, 2, \ldots, N \):
\[
\frac{dp_n(t)}{dt} = -\lambda_n p_n(t) + \lambda_{n-1} p_{n-1}(t), \text{ with } p_n(0) = 0.
\]

Recalling that we can set \( t = 1 \), the solution can be expressed in matrix notation as
\[
[p_0(1)p_1(1)\ldots p_N(1)] = [1 0 \ldots 0] \exp\{Q\}
\]
where:
\[
Q = \begin{pmatrix}
-\lambda_0 & \lambda_0 & 0 & \cdots & 0 & 0 \\
0 & -\lambda_1 & \lambda_1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & -\lambda_{N-1} & \lambda_{N-1} \\
0 & 0 & 0 & \cdots & 0 & -\lambda_N
\end{pmatrix}
\]

The use of MATLAB software enables these calculations to be done numerically.

**Temporal extension to incorporate time explicitly.**

The probabilities for a general end-point at time \( t \) take the form
\[
[p_0(t)p_1(t)\ldots p_n(t)] = [1 0 \ldots 0] \exp\{Qt\}.
\]

An obvious extension is to model two consecutive sets of data, with one sequence of \( \lambda \)'s describing the first period and a second sequence the second; i.e., if the first period is of length \( t_1 \) and the second of length \( t_2 \), then if the \( \lambda \)'s for the two periods make up matrices \( Q_1 \) and \( Q_2 \) respectively, the probabilities describing the responses at the end of the two time periods are:
\[
[p_0(t_1)p_1(t_1)\ldots p_n(t_1)] = [1 0 \ldots 0] \exp\{Q_1 t_1\}.
\]

and
\[
[p_0(t_1 + t_2)p_1(t_1 + t_2)\ldots p_n(t_1 + t_2)] = [p_0(t_1)p_1(t_1)\ldots p_n(t_1)] \exp\{Q_2 t_2\} = [1 0 \ldots 0] \exp\{Q_1 t_1\} \exp\{Q_2 t_2\}.
\]

This result can be extended to several consecutive time periods, although at the expense of more numerical calculations (with many matrix exponentiations of large, sparse matrices).

### 4 Results

Toussidou et al. (2006) advocated the use of a simpler logistic model to predict the joint impact of time and availability, simply ‘averaging’ the over-dispersion inherent in the data. Their fitting of simple generalised linear models to the individual time assays gave deviances of 296.7, 165.5, 59.5, 126.3 and 85.0 all on 18 d.f. (excepting the 24 h data which have 19
d.f.) indicating different levels of over-dispersion. Although the data are not formally replicated it is apparent that there is heterogeneity both within and between assays.

Before developing some general models incorporating time, we compared various counting process models for the 2 h foraging time. It is convenient to consider the rates $\lambda_n$ as being the product of two functions, $a_n$ and $b_{N-n}$, where $a_n$ is the generalised sequence from above, and $b_{N-n}$ is a moderating term that forces $\lambda_n$ to zero as $n \rightarrow N$. Attempts at fitting exchangeable counting process models to the data, with $b_{N-n} = (N-n)$, gave log-likelihoods markedly poorer than for the glm (log-likelihood = -71.593), so various forms of non-exchangeability were attempted. In Fenlon and Faddy (2006) we modelled $a_n$ as a simple linear function in $n$ with $b_{N-n}$ taking the form $(1-e^{-\delta(N-n)}/\delta$ or $(1-e^{-\delta(N-n)^2})/\delta$. The extra parameter $\epsilon$ provides some flexibility in the form of $b_{N-n}$. In addition to these models, an exponential form for $a_n$ was also used; it was further realised that the dependency of the number parasitised on $N$ for large $N$, caused by the limited foraging time, might be catered for by ‘forcing’ $\lambda_n$ towards zero for moderate $n$ by using a quadratic function.

Table 1 shows the potential combinations of $a_n$ and $b_{N-n}$ that were examined, and in Table 2 are shown some of the better fitting models for the 2 h data. It has already been noted that the binomial glm model is surprisingly good, though it was bettered by most of the counting process models. Interestingly the models based on B showed no dependency on $n$, and so degenerated to the A-form. The dependency returned for Model C3, but this was not as good as the two models with the exponential quadratic, D3 and D4. Minus twice the difference of log-likelihoods between these two models was only 1.2330 on 1 d.f., giving the edge to Model D4. Indeed, this was generally the best-fitting common form for $\lambda_n$ for each time assay separately, i.e.

$$\lambda_n = \exp(a + bn + cn^2) \times (N-n)^\epsilon,$$

where the quadratic term in the first component caters for time dependency in short-duration assays but also allows for possible limits to parasitoid oviposition.

Table 3 shows the log-likelihoods for Model D4 fitted to the data for each of the foraging times separately. Note that, except for the 2h data, these
represent huge improvements on the simple binomial glm models. Fitted means ± one standard deviation are shown in Figure 2. One interesting feature of the plots is the relatively constrained nature of the standard deviations which appear to be approximately constant for values of \( N > 200 \). The Figure does not indicate the shape of the distribution at particular values of \( N \). For larger values the distribution is quite symmetrical, but for small \( N \) it is highly skewed, reflecting the tendency of the response \( n \) to be close to \( N \) for the longer foraging times.

It is instructive to present the sequences for the full set of foraging times as three-dimensional bar charts, with the strips representing the sequence of \( \lambda_n \)'s for each foraging time in sequence. For clarity, four plots (showing representative values of \( N = 50, 100, 200 \) and \( 400 \)) are shown with foraging times ranged ordinally from right to left and \( n \) moving from 0 to \( N \) from foreground to background; this enables the shape of the profiles to cascade forward in a reasonably decipherable form. The sequences for Model D4 presented in Figure 3 are complex but illustrate the obvious feature of the data: that for short foraging periods the level of parasitisation is time-limited. This is indicated by the noticeably ‘peaked’ shape for short foraging times, which is effectively preserved across all values of \( N \). Note that these sequence of \( \lambda_n \)'s have not been adjusted for foraging time, so that the individual sequences relate to a specific foraging time, e.g. the first two
FIGURE 2. Fitted values ± one standard deviation of the counting process model (Model D4) for each of the five foraging times: ½, 1, 2, 5 and 24 h - each assay fitted separately.

(0-1 and 1-2) relate to the first half hour and the first hour respectively. Thus, $\lambda_0$ increases with foraging time which corresponds with the simple fact that although parasitisation may be slow to start, given sufficient time most parasitoids find and oviposit in a good proportion of aphids.

The final row of Table 3 gives the log-likelihoods under the temporal (or transition) model for the five data sets in which components of the log-likelihood for a specific foraging period are contingent on the parameters for all preceding periods. Initially the same parametric form was used for the piece-wise rates, though it was found that for $t > 1$ h $\varepsilon$ could be set to 1, implying a return to (binomial-like) exchangeability after the somewhat frenetic activity of the parasitoids during the first hour. Not surprisingly, the log-likelihood component from the first half hour’s data under the temporal model is not too different from that from the individual modelling, but the fits at 1 and 2 h are somewhat poorer; however, the components of the log-likelihood at 5 h and, particularly, 24 h represent improvements...
FIGURE 3. Sequences of $\lambda_n$ for values of $n \leq N$, where $N = 50, 100, 200$ and $400$ for foraging times of $\frac{1}{2}$, 1, 2, 5 and 24 h. Sequences are drawn as strip bars for each value of $N$ to emphasize the changes in response over time.

over the individual fits. Plots of the fitted values under the full temporal model are not generally dissimilar to those in Figure 2. In Figure 4 we present the $\lambda_n$-sequences for four typical values of $N$, but this time the bands represent time (h) so that the cross-section represents a time-dependent step function of $\lambda_n(t)$. So, for $N = 50$, for example, it can be seen that most of the activity will have taken place in the first couple of hours of foraging, although some care is needed in the interpretation of the 5 - 24 h data, for although $\lambda_0 \approx 2$ (rate per hour), declining to $\lambda_{20} \approx 1$, the latter number still represents a mean number of about 20 further 'stings' during that final period. Note that $\lambda_0$ has little meaning for this period, as the 99.9% probability interval from the fitted mean at $t = 5$ h and $N = 50$ is $15 \leq n \leq 50$. The huge 'flange' for the first half hour corresponds to the simple fact that most parasitoids will sting a fair proportion of the aphids available; thereafter, because there are fewer left to sting, it means that the rates immediately fall for the second half hour, and again for the next hour. The 2 - 5 h rates are close to zero as
there appears to be no change between these two times for \( N = 50 \). After 5 h, as already remarked, there is a slow but steady increase across the next 19 h.

For \( N = 100 \) the response is not dissimilar although the heights of the \( \lambda_n \)'s are slightly amplified. Beyond 2 h the minimum number of parasitized hosts \( (n) \) is about 35 (lower 0.0005 probability limit) at which point \( \lambda_{35} \) (2 - 5 h) is only 0.01, whereas \( \lambda_{35} \) (5 - 24 h) is 2.14 and declining slowly, with the result that at the end of 24 h the lower 0.0005 limit for \( n \) when \( N = 100 \) is 58. For \( N > 100 \) the pattern is quite similar though it is distorted somewhat by the extended scale - the proportional response is declining with increasing \( N \) so that the area of activity is compressed and relatively closer to the origin. There are two obvious changes: for \( t = 2 \) - 5 h activity starts to increase beyond \( n = 75 \) or so, and there is an exaggerated downward ramp for \( \lambda_n \) (5 - 24 h) as \( n \) increases. The first of these seems to correspond to those parasitoids that strike at a fast rate in the first couple of hours continuing for a further few hours, whereas the second is almost
the opposite with ‘slow-starting’ parasitoids playing ‘catch-up’ in the latter part of the foraging period.

5 Conclusion

We have developed a Markov model of temporally-ordered functional response data. The original data came from a series of replicated assays in which individual parasitoids were given access to different numbers of aphids for fixed periods of time. The significant advance we present here is the extension of a basic counting process model to a full temporal model, using a time-dependent Markov model with piece-wise constant rates. The first important fact is that this provides a better fit to the data than do the individual fits for each foraging time, although the benefit is most pronounced at 24 h. What this modelling also demonstrates is the marked changes in foraging behaviour over a 24 h period: within 1 h much has been achieved, and the rate of new parasitisation in the second hour is very small even when large numbers of hosts are available. Beyond 2 h it is interesting to note that for smaller values of $N(\leq 100$, say) consolidation occurs, and the wasps seek out and parastise the majority of remaining hosts, whereas for larger values of $N$ the number of parasitized hosts continues to increase, though not as rapidly as in the first hour or so.

6 References


The reliability of censored reliability analyses

H. K. Finselbach\(^1\) and A. J. Watkins\(^1\)

\(^1\) School of Business & Economics, University of Wales Swansea, Singleton Park, Swansea, SA2 8PP, United Kingdom, a.watkins@swansea.ac.uk

**Abstract:** We consider the analysis of reliability data subject to Type I censoring, and, in particular, the extent to which an interim analysis provides a guide to the final analysis. For simplicity, we assume that the lifetimes of items follow the negative exponential distribution, and present asymptotic results on the correlation between two estimates of mean lifetime \(\theta\). This, in turn, yields 95\% confidence limits for the final estimate given the interim estimate. We illustrate our results using two published data sets, and also present results from simulation experiments which indicate the extent to which the asymptotic results apply in samples of finite size.

1 Introduction

Although data available for statistical analysis is typically regarded as fixed once collected, there are several exceptions: these include time series analysis, longitudinal studies, and reliability and survival analysis. In such cases, data generally accrues with time, and it is possible to conduct one or more interim analyses in addition to a final analysis; successive analyses are then based on increasing sample sizes. In the final case, there is the additional feature that the status of items (for instance, whether operational or not) can vary with time, even when the sample size is fixed at the outset of an experiment.

This paper presents some results which link the statistical analyses of reliability data arising from a sample of items at two or more points in time. For simplicity, we assume that the lifetimes of individual items follow the negative exponential distribution with mean \(\theta\), that a set of \(n\) such items are put on test simultaneously, and that continual monitoring of these items is possible. Under the assumption that the items yield statistically independent observations, the maximum likelihood estimator of \(\theta\) is well-known; see, for example, Nelson (1982). At any given time \(c\), say, the experiment is subject to Type I censoring, in that we only have the exact lifetimes \(Y_1, \ldots, Y_M\) of the \(M\) \((0 \leq M \leq n)\) items that have failed before \(c\), with the other \(n - M\) items having a censored operational life of \(c\). Thus, with \(M > 0\), we have

\[
\hat{\theta}_c = \frac{S_M + (n - M)c}{M}
\]  

(1)
TABLE 1. Summary of $\hat{\theta}_c$ for three values of $c$ for the electrical component data.

<table>
<thead>
<tr>
<th>$c$</th>
<th>$M$</th>
<th>$S_M$</th>
<th>$\hat{\theta}_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>5</td>
<td>48</td>
<td>34.6</td>
</tr>
<tr>
<td>50</td>
<td>8</td>
<td>152</td>
<td>31.5</td>
</tr>
<tr>
<td>$\infty$</td>
<td>10</td>
<td>218</td>
<td>28.8</td>
</tr>
</tbody>
</table>

where $S_M = \sum_{i=1}^{M} Y_i$; this estimate of $\theta$ yields, in turn, estimates of percentiles of item life. In practice, it may be possible to repeat the above analysis at each of a sequence $c_1, c_2, \ldots$ of times, until (at a sufficiently large value for $c$) all items have failed and the data set is complete. From the complete data, we obtain $\hat{\theta}_\infty \equiv \hat{\theta}$, while $\hat{\theta}_c \to \theta$ as $c \to \infty$.

1.1 Example

We first use data from Kalbfleisch (1979) to illustrate the above experimental set-up. The failure time (in days) of $n = 10$ electrical components are

<table>
<thead>
<tr>
<th></th>
<th>70</th>
<th>11</th>
<th>66</th>
<th>5</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
<td>35</td>
<td>40</td>
<td>29</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 1 shows the details behind the calculation of $\hat{\theta}_c$ for $c = 25, 50$ and $\infty$; more generally, we can plot $\hat{\theta}_c$ against $c$. Similar calculations are possible for the $n = 49$ failure times below, taken from Epstein (1960); the discussion there means we may regard the negative exponential to be a reasonable model for this data.

<table>
<thead>
<tr>
<th></th>
<th>1.2</th>
<th>7.0</th>
<th>23.9</th>
<th>47.9</th>
<th>62.7</th>
<th>95.1</th>
<th>128.7</th>
<th>151.6</th>
<th>185.2</th>
<th>253.1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.2</td>
<td>12.1</td>
<td>24.3</td>
<td>48.4</td>
<td>72.4</td>
<td>97.9</td>
<td>133.6</td>
<td>152.6</td>
<td>187.1</td>
<td>304.1</td>
</tr>
<tr>
<td></td>
<td>4.9</td>
<td>13.7</td>
<td>25.1</td>
<td>49.3</td>
<td>73.6</td>
<td>99.6</td>
<td>144.1</td>
<td>164.2</td>
<td>203.0</td>
<td>341.7</td>
</tr>
<tr>
<td></td>
<td>5.0</td>
<td>15.1</td>
<td>35.8</td>
<td>53.2</td>
<td>76.8</td>
<td>102.8</td>
<td>147.6</td>
<td>166.8</td>
<td>204.3</td>
<td>354.4</td>
</tr>
<tr>
<td></td>
<td>6.8</td>
<td>15.2</td>
<td>38.9</td>
<td>55.6</td>
<td>83.8</td>
<td>108.5</td>
<td>150.6</td>
<td>178.6</td>
<td>229.5</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 gives a summary of $\hat{\theta}_c$ calculated at several censoring times for these $n = 49$ failure times, while Figure 1 displays both data and the relationship between $\hat{\theta}_c$ and $c$, in which we observe sharp drops in $\hat{\theta}_c$ at the actual times to failure - reflecting the increment in $M$, the denominator of $\hat{\theta}_c$ - offset by linear increases in $\hat{\theta}_c$ between failure times.

From a statistical perspective, the analysis of the complete data set is to be preferred, while censoring (though undesirable from a statistical perspective) is often necessary in terms of the timescale for the experiment, and costs involved in running the experiment. From a practical perspective, we are interested in gauging the earliest point in time $c$ at which the experiment can be reasonably terminated with $\hat{\theta}_c$ still yielding a close or reliable
TABLE 2. $\hat{\theta}_c$ for various $c$ for $n = 49$ failure times.

<table>
<thead>
<tr>
<th>$c$</th>
<th>25</th>
<th>75</th>
<th>125</th>
<th>175</th>
<th>225</th>
<th>275</th>
<th>325</th>
<th>$\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\theta}_c$</td>
<td>88.03</td>
<td>114.97</td>
<td>124.46</td>
<td>114.07</td>
<td>108.68</td>
<td>107.92</td>
<td>108.37</td>
<td>104.89</td>
</tr>
</tbody>
</table>

FIGURE 1. The $n = 49$ failure times ($\times$) from Epstein (1960), and $\hat{\theta}_c$ versus $c$ for this data set.

guide to $\hat{\theta}$ - this explains our title. In the following sections, we consider the relationship between $\hat{\theta}$ and $\hat{\theta}_c$, and hence the extent to which $\hat{\theta}_c$ can be regarded as a reliable guide to $\hat{\theta}$.

2 Link between $\hat{\theta}$ and $\hat{\theta}_c$

From (1), we have

$$n\hat{\theta} = M\hat{\theta}_c + \sum_{i=M+1}^{n} (Y_i - c)$$
where $Y_{M+1}, \ldots, Y_n$ denote the lifetimes of items still operational at $c$, and $M$ follows a Binomial distribution with parameters $n$ and $q_c$ given by

$$q_c = 1 - \exp\left(-\frac{c}{\theta}\right)$$

which is the probability that any item fails in $(0, c)$. Moments of $M$ appear in the expected Fisher information for $\hat{\theta}_c$, which is $nq_c\theta^{-2}$; it is also possible to show that the usual asymptotic relationship between the maximum likelihood estimator, the expected Fisher information and the score function is

$$\sqrt{nq_c\theta} \left(\hat{\theta}_c - \theta\right) \simeq \sqrt{\frac{\theta^2}{nq_c}} \left[-M\theta^{-1} + \theta^{-2} \{S_M + (n - M)c\}\right].$$

Since this relationship also covers the case $c \to \infty$ (when it becomes exact), we now approximate

$$\text{Corr} \left\{\hat{\theta}_c, \hat{\theta}_c\right\} = \text{Corr} \left\{\sqrt{nq_c\theta} \left(\hat{\theta}_c - \theta\right), \sqrt{\frac{\theta^2}{nq_c}} \left(\hat{\theta}_c - \theta\right)\right\}$$

by

$$\text{Corr} \left\{\sqrt{\frac{\theta^2}{n}} \left[-n\theta^{-1} + \theta^{-2}S_n\right], \sqrt{\frac{\theta^2}{nq_c}} \left[-M\theta^{-1} + \theta^{-2} \{S_M + (n - M)c\}\right]\right\}$$

We can now use regularity conditions to write this last correlation in terms of expectations, and hence obtain

$$\frac{\theta^2}{n\sqrt{q_c}} E \left[\left[-n\theta^{-1} + \theta^{-2}S_n\right] \times \left[-M\theta^{-1} + \theta^{-2} \{S_M + (n - M)c\}\right]\right] \quad (2)$$

as our approximation to $\text{Corr} \left\{\hat{\theta}_c, \hat{\theta}_c\right\}$.

### 2.1 Some statistical considerations

We now consider terms emerging from the expansion of (2). Since $Y_1, \ldots, Y_M$ follow the right truncated negative exponential distribution with probability density function $\theta^{-1}q_c^{-1} \exp\left(-\theta^{-1}y\right)$ for $0 \leq y \leq c$, we have from Watkins and John (2004)

$$E [Y_i] = \theta - cr_c, \quad (3)$$

and

$$E [Y^2_i] = 2\theta^2 - 2\theta cr_c - c^2 r_c, \quad (4)$$
for $i = 1, \ldots, M$, where
\[ r_c = \frac{1 - q_c}{q_c} \]
is the odds ratio of survival beyond $c$ to failure before $c$. Thus,
\[ E \left[ n\theta^{-2}M \right] = n^2\theta^{-2}q_c \]
cancels with
\[ E \left[ -n\theta^{-3} \{ S_M + (n - M) c \} \right] = -n^2\theta^{-2}q_c, \]
so that our approximation to $Corr \left\{ \hat{\theta}, \hat{\theta}_c \right\}$ reduces to
\[ \frac{1}{n\sqrt{q_c}} E \left[ S_n \times \left\{ -M\theta^{-1} + \theta^{-2} \{ S_M + (n - M) c \} \right\} \right]. \] (5)
Since $M, S_M$ and $S_n$ are inter-related, the expectations $E [MS_n]$ and $E [S_nS_M]$ needed from (5) require some care, but can be obtained using
\[ S_n = S_M + \sum_{i=M+1}^{n} Y_i \]
in which the mean lifetime of items $Y_{M+1}, \ldots, Y_n$ censored at $c$ is, via the lack-of-memory property of the negative exponential distribution, given by $c + \theta$. Thus, we have
\[ E [MS_n] = E \left[ MS_n + M \sum_{i=M+1}^{n} Y_i \right] = \{ \theta - cr_c \} E \left[ M^2 \right] + (c + \theta) E \left[ nM - M^2 \right], \]
via (3) and a standard conditioning argument; substituting the moments of $M$ and simplifying gives
\[ E [MS_n] = n^2\theta q_c - nc (1 - q_c). \]
Likewise, from (3) and (4) and a similar conditioning argument, we have
\[ E [S_nS_M] = E \left[ \left( S_M + \sum_{i=M+1}^{n} Y_i \right) S_M \right] = (2\theta^2 - 2\theta c r_c - c^2 r_c) E \left[ M^2 \right] + (\theta - cr_c)^2 E \left[ M^2 - M \right] + (c + \theta) (\theta - cr_c) E \left[ nM - M^2 \right]; \]
substituting the moments of $M$ and simplifying gives
\[ E [S_nS_M] = n\theta^2 q_c - n\theta c (1 - q_c) + n^2\theta^2 q_c - n^2\theta c (1 - q_c) - nc^2 (1 - q_c). \]
TABLE 3. $\text{Corr} \left( \hat{\theta}, \hat{\theta}_c \right)$ for negative exponential data generated with $\theta = 100$. Figures are based on 10000 replications.

<table>
<thead>
<tr>
<th>$c$</th>
<th>$\sqrt{q_c}$</th>
<th>$n = 25$</th>
<th>$n = 100$</th>
<th>$n = 1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.6273</td>
<td>0.5359</td>
<td>0.6072</td>
<td>0.6313</td>
</tr>
<tr>
<td>150</td>
<td>0.8814</td>
<td>0.8697</td>
<td>0.8787</td>
<td>0.8815</td>
</tr>
<tr>
<td>250</td>
<td>0.9581</td>
<td>0.9534</td>
<td>0.9578</td>
<td>0.9576</td>
</tr>
</tbody>
</table>

Finally, substituting these expressions for $E\left[MS_n\right]$ and $E\left[S_nS_M\right]$ into (5), we obtain, after further simplification,

$$\text{Corr} \left( \hat{\theta}, \hat{\theta}_c \right) \simeq \sqrt{q_c}.$$  

Since this result is an asymptotic one, we next consider the extent to which it applies in finite samples. Table 3 summarises some simulation experiments, and we observe good agreement between theory and practice with $n, c$ increasing.

Furthermore, the asymptotic Normality of maximum likelihood estimators (Cox & Hinkley, 1974) means that the asymptotic distribution of $\hat{\theta} - \hat{\theta}_c$ is also Normal, with zero mean and variance

$$\frac{\theta^2 r_c}{n},$$

so that a 95% confidence interval for $\hat{\theta}$ based on $\hat{\theta}_c$ is

$$\hat{\theta} = \hat{\theta}_c \pm 1.96\hat{\theta}_c \sqrt{\frac{r_c}{n}}.$$  

Figure 2 shows these limits for various $c$ for the $n = 49$ times to failure from Epstein (1960). As above, we are also interested in the extent to which these limits apply in finite samples; Table 4 provides some summaries of simulation experiments to assess for given $n, c$, and, again, we see generally good agreement between theory and practice.

3 Generalisations

We can easily extend the above analysis to the case where data is subject to Type II Censoring. Another obvious extension is to the Weibull distribution, where we are not only interested in the two parameters $\beta, \theta$ in the distribution, but also in the quantile function.
FIGURE 2. $\hat{\theta}_c$ and 95% confidence limits for $\hat{\theta}$ for various $c$ for the $n = 49$ times to failure (×) from Epstein (1960).
TABLE 4. % of $\hat{\theta}$ in the 95% Confidence Interval when fitted to negative exponential data generated with $\theta = 100$. Figures are based on 10000 replications.

<table>
<thead>
<tr>
<th>c</th>
<th>$n = 25$</th>
<th>$n = 100$</th>
<th>$n = 1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>94.78</td>
<td>94.92</td>
<td>95.31</td>
</tr>
<tr>
<td>150</td>
<td>95.23</td>
<td>95.13</td>
<td>94.97</td>
</tr>
<tr>
<td>250</td>
<td>95.69</td>
<td>95.24</td>
<td>95.00</td>
</tr>
</tbody>
</table>

$B_p = \theta \left\{-\ln(1-p)\right\}^{\frac{1}{\beta}}$

for specified $p$. The analysis here will involve the same asymptotic relationships between the maximum likelihood estimators of the two parameters and the score functions and the expected Fisher information matrices of the likelihoods, but the extra parameter makes the analysis more involved. Furthermore, the form of $B_p$ means that the correlation between its estimates at $c$ and $\infty$ must be based on a first order Taylor series expansion of $B_p$; this introduces a further approximation which, in turn, requires further checks in the assessment of the extent to which asymptotic results apply in finite samples. The overall analysis is thus considerably longer than that given here, and will therefore be presented elsewhere.

References


Generalized nonlinear models, and over-parameterized representation

David Firth$^1$ and Heather L. Turner$^1$

$^1$ Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom, d.firth@warwick.ac.uk

Abstract: This paper will describe and illustrate various aspects of the design and implementation of the gnm package (Turner and Firth, 2005) for the R statistical computing environment (R Development Core Team, 2005). The gnm package facilitates the flexible specification, fitting and inspection of generalized nonlinear models. The systematic use of over-parameterized model representations, in which the mapping from parameters to distributions is many-to-one, is emphasized. Methods are described for maximum likelihood computation and inference, and for presenting conclusions, in the presence of deliberate parameter redundancy.

Keywords: Generalized inverse; iterative weighted least squares; parameter redundancy; quasi variances.

1 Introduction

A generalized nonlinear model, for our purposes here, is a parametric regression model for univariate response $Y$, in which

- the link-transformed mean $g[E(Y)]$ is equal to a specified function $h(\beta_1, \ldots, \beta_p)$ of unknown parameters $\{\beta_r : r = 1, \ldots, p\}$
- the variance is equal to, or proportional to, a known function $v[E(Y)]$.

This class includes all generalized linear models, as well as models in which $h$ is not linear in the unknown parameters. It also includes parametric least-squares nonlinear regressions, as the special case where $v$ is a constant. Applications often, but not necessarily, involve a ‘partially linear’ predictor function $h$ which can be decomposed as

$$h(\beta_1, \ldots, \beta_p) = \sum_{r=1}^{k} x_{ir}\beta_r + h_2(\beta_{k+1}, \ldots, \beta_p),$$

with $h_2$ a non-linear function of the relevant subset of parameters. Included in this large class are row-column association models (Goodman,
D. Firth and H.L. Turner (1979), ‘GAMMI’ (generalized additive main affects and multiplicative interaction) models (e.g., van Eeuwijk, 1995), log-multiplicative models for social mobility (Xie, 1992; Erikson and Goldthorpe, 1992), Rasch-type logit or probit models for legislative voting (e.g., de Leeuw, 2006), ‘diagonal reference’ structures for dependence on a square or hyper-square classification (Sobel, 1981, 1985), ‘stereotype’ multinomial regression models for ordinal response (Anderson, 1984), and many more. Response-variable types include all of those that are familiar from generalized linear models.

The gnm package allows any generalized nonlinear model, subject to differentiability and other reasonable behaviour of the functions involved, to be fitted by maximum likelihood or quasi-likelihood, with specification via a model formula which aims to be as ‘natural’ as possible. This last requirement motivates an approach in which the parametric representation of the model — for example, the parameter constraints/conventions needed to ensure identifiability — is regarded as an arbitrary aspect and not part of the model specification. For example, a row-column association model with linear predictor of the form

\[ \alpha_r + \beta_c + \delta_r \gamma_c \]

is naturally expressed in gnm as an S-language model formula,

% y ~ row + column + Mult(row, column)

without regard for any constraints that might ultimately needed in order to make the specific symbols \( \alpha_r, \beta_c, \delta_r, \gamma_c \) meaningfully interpretable. Inference on specific, identifiable parameter combinations is regarded, instead, as being quite separate from the act of specifying the model.

This approach demands some novel technical solutions. These include: the development of an iterative weighted least squares algorithm using a generalized inverse of the local information matrix; a computationally efficient approach to handling large numbers of nuisance parameters, which generalizes the eliminate facility that was first seen in GLIM (e.g., Hatzinger and Francis, 2004); and the implementation of methods to check the identifiability of specific parameter combinations (Catchpole and Morgan, 1997) and to present estimates of certain sets of parameter combinations, especially contrasts, in a maximally informative way (Firth, 2003; Firth and Menezes, 2004).

A rapid overview will be given of all of the above, with illustrative examples derived from real applications in social science and agriculture. Further examples and much more detail can be found in Turner and Firth (2005).

Acknowledgments: This work was supported by the Economic and Social Research Council (United Kingdom) through a Professorial Fellowship.
References


A longitudinal transition ordered response model for analyzing the change of attitudes to gender roles

M. Ganjali¹, D. Berridge² and R. Penn²

¹ Department of Statistics, Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran, m-ganjali@sbu.ac.ir
² Centre for Applied Statistics, Department of Mathematics and Statistics and Department of Sociology, Fylde College, Lancaster University, Lancaster LA1 4YF, UK, d.berridge@lancaster.ac.uk and r.penn@lancaster.ac.uk

Abstract: This paper examines changes in attitudes to gender roles in contemporary Britain by using a first order Markov process and a cumulative logistic approach. The data are taken from the British Household Panel Study (BHPS). The attitudinal responses examined take the form of ordinal responses concerning gender roles in 1991, 1997 and 2003. The conditional likelihood function is partitioned to make possible the use of existing software for estimating model parameters. For the BHPS data, we found that, depending on the value of the response in 1991 (1997), different factors are important determinants of attitudes to gender roles in 1997 (2003).

Keywords: Longitudinal ordinal data; transition model; cumulative logit model; conditional likelihood; gender roles.

1 Introduction

The issues of gender, economic activity and the domestic division of household tasks have been of major interest to sociologists in Britain over the last thirty years. Roberts (1985) provided a picture of classic traditionalistic gender roles in her study of the working class in North West England during the inter-war years. Women were expected to get married by their mid-twenties and to withdraw from paid employment outside the home in order to concentrate on domestic work and child rearing. Indeed, the 1951 Census of Population showed that only 26 percent of married women were economically active (Witherspoon, 1988).

This pattern of unequal and segregated gender roles displayed remarkable persistence in postwar Britain (Hakim, 1995). Indeed, Roberts (1995), in an oral history of the period between 1940 and 1870, found strong evidence of continued asymmetric gender roles. Successive British Social Attitudes Surveys have examined the form and determinants of attitudes to the
domestic allocation of household tasks between men and women. Wither- 
spoon’s analysis reconfirmed that the bulk of household work was still done 
by women (Witherspoon, 1988). She reported also that more highly quali- 
fied women were more likely to hold egalitarian views. Kiernan (1992) also 
found little evidence of change. She reported that attitudes to gender roles 
varied by age, level of education and employment status. There is little ev- 
idence of fundamental change since then (Abercrombie and Warde, 1992; 
Scott et al., 1998.)

On the basis of the existing literature, it is clear that we can broadly distin-
guish two types of view about gender roles: a traditionalistic, segregationist, 
inegalitarian type and a non-traditionalistic, shared, egalitarian one. ‘Tra- 
ditionalists’ believe in segregated gender roles wherein women should look 
after children at home and undertake most domestic activities whilst men 
should provide the financial support for the family in their roles as ‘bread- 
winner’ and ‘head of household’. The ‘egalitarian’ or ‘symmetrical’ image 
of gender roles involves the assumption that such inegalitarian patterns are 
in the process of breaking down.

Nonetheless, a range of empirical factors have been identified that affect the 
likelihood that an individual would fit this ‘traditionalistic’ model. Age has 
been shown to be significant: older people are more likely to be ‘traditional- 
istic’ than younger people. Gender, unsurprisingly, has been shown repeat-
edly to be of major importance: women are less likely than men to hold 
‘traditionalistic’ orientations. Household size has been another important 
factor: women with large families are generally locked into a traditionalistic 
pattern of gender roles (Ahmad, 1998). The employment status of women 
has also been shown to be highly significant. Full-time female employees 
are more likely to reject ‘traditionalistic’ orientations. It appears also that 
‘traditionalism’ is inversely related to level of qualifications, those with the 
highest levels of qualifications being the least ‘traditionalistic’.

However, the ordinal nature of attitudinal response to gender roles is not 
enscraped systematically within any of the correlational analyses out-
lined in the earlier review of the literature. Indeed, they generally utilize a 
simplified dichotomous response (‘agree’ versus ‘not agree’), thereby losing 
a considerable amount of information contained within the response items. A variety of ordinal regression models can be used to model such responses 
(Agresti, 2002). These include cumulative models (with logit, probit or com-
plementary log-log link functions) and sequential models (Berridge, 1995, 
Tutz, 2005).

In a panel or longitudinal study, where there is a sequence of ordinal re-
sponses for the same individual, we have to consider not only the fact that 
responses are ordinal in nature but also the possibility of correlations be-
tween responses given by the same individual. Different models can be used 
to handle such correlations. One possibility is marginal modelling, which 
can be used to make inferences about population averaged parameters (Ten 
Have et al., 1996; Kim, 1995). A second possibility is random effects mod-
M. Ganjali et al. 195

elling, which makes inferences about variability between respondents (Tutz and Hennevogl, 1996; Verbeke and Molenberghs, 1997). For the data considered here the primary question of interest is how attitudes to gender roles are changing over time or how transitions are made between consecutive time points. For such a scientific question, a more appropriate approach would be to use Markov (transition) models (see Garber, 1989, for using multinomial logit). For reviews of transition and other models for ordinal longitudinal responses, see Diggle et al. (2002) and Agresti (1999, 2002).

In this paper, we use a first order Markov transition model for repeated ordinal responses to assess the relative significance of a range of explanatory variables concerning attitudes to gender roles over time. This will be done by partitioning the conditional likelihood for different values of the previous response and by using existing software.

In Section 2, the British Household Panel Study (BHPS) data are introduced. In Section 3, transition model and the conditional likelihood are presented. In Section 4, results of applying the transition model to the BHPS data are discussed.

2 The data

The data have been extracted from the British Household Panel Survey (BHPS): a longitudinal survey that has been conducted annually since 1991. Information was collected through personal interviews. Every other year, the BHPS also includes a ‘Living in Britain’ survey, in which respondents are asked to complete in person a confidential questionnaire. The questions asked in this survey cover a wide range of subjects. One section of the questionnaire asks the respondents to say whether they personally agree or disagree with a number of statements related to gender roles. One of these statements is ‘A husband’s job is to earn money; a wife’s job is to look after the home and family’. In terms of the gender roles outlined above, this proposition is a traditionalistic, segregated or non-egalitarian orientated question. To answer the question, respondents are asked to choose one of the following ordinal categories: SA: Strongly agree; A: Agree; NAND: Neither agree nor disagree; D: Disagree and SD: Strongly disagree. The responses examined in this analysis are taken from wave 1 (1991), wave 7 (1997) and wave 13 (2003). These three waves were chosen to give us the best opportunity to observe changes in attitudes towards gender roles from wave 1 (baseline) to wave 7 and from wave 7 to wave 13. In 1991, 10264, in 1997, 11193 and in 2003, 16238 respondents were included in the sample. This clearly represents a very complicated data set with a high level dropout at the individual level and many boosters into the BHPS. The sample involved 5474 respondents who answered the question in all three waves. Most variables were not available for the respondents not included in the current sample. Such lack of information would severely handicap
TABLE 1. Level of agreement with the proposition ‘A husband’s job is to earn money; a wife’s job is to look after the home and family’ (1991 v. 1997 and 1997 v. 2003, observed counts and row percentages)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SA</td>
<td>A</td>
<td>NAND</td>
</tr>
<tr>
<td>1991</td>
<td>86</td>
<td>79</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>32.6%</td>
<td>29.9%</td>
<td>17.8%</td>
</tr>
<tr>
<td>A</td>
<td>63</td>
<td>262</td>
<td>290</td>
</tr>
<tr>
<td></td>
<td>9.3%</td>
<td>38.5%</td>
<td>33.8%</td>
</tr>
<tr>
<td>NAND</td>
<td>33</td>
<td>138</td>
<td>565</td>
</tr>
<tr>
<td></td>
<td>3.2%</td>
<td>13.4%</td>
<td>48.9%</td>
</tr>
<tr>
<td>D</td>
<td>19</td>
<td>98</td>
<td>351</td>
</tr>
<tr>
<td></td>
<td>1.3%</td>
<td>6.5%</td>
<td>23.3%</td>
</tr>
<tr>
<td>SD</td>
<td>5</td>
<td>13</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>0.6%</td>
<td>1.6%</td>
<td>10.4%</td>
</tr>
<tr>
<td>Total</td>
<td>206</td>
<td>590</td>
<td>1216</td>
</tr>
<tr>
<td></td>
<td>4.8%</td>
<td>13.8%</td>
<td>28.4%</td>
</tr>
</tbody>
</table>

The association parameter (gamma, Agresti, 2002) between responses in 1991 v. 1997 takes the value 0.652 (s.e.=0.013) and between responses in 1997 v. 2003 takes the value 0.728 (s.e.=0.011).
3 Ordered transition model and conditional likelihood

For a univariate response, the cumulative logit model estimates the effects of explanatory variables on the log odds of selecting lower, rather than higher, response categories. This model is

$$\log\left( \frac{\text{pr}(Y_i \leq b; \alpha, \beta)}{\text{pr}(Y_i > b; \alpha, \beta)} \right) = \alpha_b - \sum_{k=1}^{K} \beta_k X_{ik} \text{ for } b = 1, \ldots, J - 1$$

(1)

In the above equation, $Y_i$ is the response of the $i$th individual, $X_{ik}$ is the $k$th explanatory variable of the $i$th individual, $J$ is the number of response categories, $\alpha_b$s are intercepts (cut-points) indicating the logarithms of odds of selecting lower rather than higher categories when all explanatory variables equal zero, $\alpha = (\alpha_1, \ldots, \alpha_{J-1})$ is the vector of cut-point parameters, $\beta = (\beta_1, \ldots, \beta_K)$ is the vector of regression coefficients related to explanatory variables and $K$ is the number of explanatory variables. A positive coefficient indicates increased likelihood of selecting a higher response category. The cumulative logit model assumes that the effects of different explanatory variables are fixed across all $J - 1$ partitions of the ordinal response. McCullagh (1980) demonstrated that such a model possesses the properties of a proportional odds model. This can be implemented readily in software such as SPSS (ordinal regression) or STATA (ordered logit regression).

In the following we shall use a first-order transition model based on the cumulative logit framework and the conditional likelihood i.e. we shall condition on the first response, $Y_{i1}$, for each respondent. The form of the transition model for $t = 2, 3, a = 1, \ldots, J - 1$ and $b = 1, \ldots, J - 1$ is:

$$\log\left( \frac{\text{pr}(Y_{it} \leq b|Y_{it-1} = a; \alpha_{at}, \beta_{at})}{\text{pr}(Y_{it} > b|Y_{it-1} = a; \alpha_{at}, \beta_{at})} \right) = \alpha_{atb} - \sum_{k=1}^{K} \beta_{atk} X_{itk}.$$  

(2)

where $Y_{i1}, Y_{i2}$ and $Y_{i3}$ are the responses given by the $i$-th individual in 1991, 1997 and 2003 respectively, $\alpha_{at} = (\alpha_{at1}, \ldots, \alpha_{at(J-1)})$ is the vector of cut-point parameters for $a = 1, \ldots, J$ and $t = 2, 3$, and $\beta_{at} = (\beta_{at1}, \ldots, \beta_{atK})$ is the vector of parameters related to explanatory variables for $a = 1, \ldots, J$ and $t=2,3$. Explanatory variables are permitted to be dynamic i.e. vary over time.

Denote the conditional probabilities $\text{pr}(Y_{t} = b|Y_{t-1} = a; \alpha_{at}, \beta_{at})$ by $\Pi_{b|a}(t)$ which are called transition probabilities. The $J \times J$ matrix $\{\Pi_{b|a}(t), a = 1, \ldots, j; b = 1, \ldots, J\}$ is a first order transition probability matrix. Let $n_{ab}(t)$ denote the number of transitions from state $a$ at time $t - 1$ to state $b$ at time $t$. If subjects behave independently the conditional likelihood function is:

$$L = \prod_{t=2}^{3} \prod_{a=1}^{J} \prod_{b=1}^{J} \left[ \Pi_{b|a}(t) \right]^{n_{ab}(t)}$$

(3)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Sig. in 1997</th>
<th>Sig. in 2003</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA</td>
<td>Age (+), Male (+), Widowed (-), Self employed (-), Full- or Part-time employed (-), Home maker (+)</td>
<td>SA</td>
</tr>
<tr>
<td>A</td>
<td>Age (+), Male (+), Higher or first degree (-), 4 people in house (-)</td>
<td>A</td>
</tr>
<tr>
<td>NAND</td>
<td>Age (+), Male (+), Other higher (-)</td>
<td>NAND</td>
</tr>
<tr>
<td>D</td>
<td>Age (+), Male (+), Home maker (+), Higher or first degree (-), Other Higher degree (-), Technicians (-), Labour supporter (-)</td>
<td>D</td>
</tr>
<tr>
<td>SD</td>
<td>Male (+), Home maker (+), Owned the house (+), Higher or first degree (-)</td>
<td>SD</td>
</tr>
</tbody>
</table>

where for $t=2,3$

$$\Pi_{b|a(t)} = pr(Y_{it} \leq b|Y_{it-1} = a; \alpha_{at}, \beta_{at}) - pr(Y_{it} \leq b-1|Y_{it-1} = a; \alpha_{at}, \beta_{at}),$$

and

$$pr(Y_{it} \leq b|Y_{it-1} = a; \alpha_{at}, \beta_{at}) = \frac{\exp(\alpha_{ath} - \sum_{k=1}^{K} \beta_{atk}X_{itk})}{1 + \exp(\alpha_{ath} - \sum_{k=1}^{K} \beta_{atk}X_{itk})}, \quad (4)$$

Assume that the effects of explanatory variables observed in 1997 (2003), on the probability of response in 1997 (2003), are constant, regardless of the observed value of the response in 1991 (1997). Under this assumption, likelihood (3) can be maximized using any appropriate optimization method such as Newton-Raphson or Quasi-Newton-Raphson. When we relax this assumption and allow the effects of explanatory variables in 1997 (2003), on 1997 (2003) response probabilities, to vary with different values of the response in 1991 (1997), the maximization of likelihood (3) may be performed using existing software such as SPSS or STATA.

4 Results

Table 2 lists those explanatory variables (see appendix) which were significant determinants of attitudes toward gender roles in 1997 (2003) given the...
attitudes of respondents in 1991 (1997). Positive and negative signs correspond to the effects on the cumulative probability of the response being ‘in agreement’. For example, gender had a consistently significant effect on responses in 1997, regardless of attitudes in 1991. In the results for 1997, males were more likely than females to agree. Respondents widowed by 1997 were significantly more likely to disagree in 1997, having said SA in 1991. Age was a significant determinant of attitudes in 2003 for those respondents who said A, NAND or D in 1997. The older the respondent, the more likely he/she was to agree in 2003. The other significant effects can be interpreted in a similar manner.

References


1. Gender (1=male, 2=female)
3. Number of people in household (1, 2, 3, 4, 5 or more)
4. Marital status (1=married; 2=living as a couple; 3=widowed; 4=divorced or separated; 5=never married)
5. Housing tenure (1=owned outright; 2=owned with mortgage; 3=rented)
6. Current economy activity (1=self employed; 2=full time employee; 3=part time employee; 4=unemployed; 5=homemaker; 6=other)
7. Highest educational qualification (1=higher or first degree; 2=other higher qualifications; 3=other qualifications; 4=no qualifications)
8. Lancaster social class: present job (1=coordinator; 2=entrepreneurs; 3=clerical; 4=skilled manual or technicians; 5=not applicable; 6=non-skilled)
9. Political party supported (1=Conservative; 2=Labour; 3=Liberal Democrat or Nationalist or Social Democratic Party; 4=other)
Additive mixed models with P-splines

Sonja Greven\textsuperscript{1,2}, Helmut Küchenhoff\textsuperscript{1} and Annette Peters\textsuperscript{2} for the AIRGENE study group

\textsuperscript{1} LMU, Department of Statistics, Akademiestr. 1, D-80799 Munich, Germany, sonja.greven@stat.uni-muenchen.de
\textsuperscript{2} GSF National Research Center for Environment and Health, Institute of Epidemiology, Ingolstädter Landstrasse 1, D-85764 Neuherberg, Germany

1 Overview

We present an approach to additive mixed models using P-Splines where the spline coefficients are estimated in the mixed model framework. We have implemented this approach in a SAS macro which can be used in a wide variety of contexts. We apply our methodology to an epidemiological study assessing the health effects of ambient air pollution.

2 Motivating data set - the AIRGENE Study

Background: Epidemiological studies clearly link ambient air pollution, especially particulate matter (PM), to morbidity and mortality due to cardio-pulmonary diseases. This lead to new regulations being introduced in the EU in 2005, limiting average PM\textsubscript{10} (\(\bar{0} < 10 \mu m\)) concentrations in cities. However, research on causal pathways linking air pollution to outcomes such as myocardial infarctions is still ongoing. Pro-inflammatory and pro-thrombotic processes are thought to be involved. There might also be genetic dispositions for susceptibility to air pollutants.

The study: The AIRGENE study is an EU-funded epidemiological study conducted in the six European cities Athens, Augsburg, Barcelona, Helsinki, Rome and Stockholm between May 03 and July 04. It aims at assessing inflammatory responses in association with ambient air pollution concentrations in myocardial infarction (MI) survivors (see also Ruckerl et al., 2005) and at defining susceptible subgroups of MI survivors based on genotyping.

Data structure: Three inflammatory blood markers (C-reactive protein, Fibrinogen and Interleukin-6) were measured every month repeatedly up to 8 times in over 1,000 MI survivors, resulting in about 6,000 samples per marker in total. Ten air pollution and several weather variables were measured hourly throughout the study period. Patient characteristics were
collected at baseline, including the determination of 114 SNPs on 13 inflammatory pathway genes hypothesized to modify the pollutant effects.

3 Methods

To analyze the AIRGENE data, we have to account for the longitudinal data structure and the potential non-linearity of weather and trend variables. We decided on a mixed model approach using a random patient effect. Short half-times of the blood markers render an additional correlation structure unnecessary. Trend and weather variables are potentially included as additive terms modelled by P-Splines, where coefficients are estimated in the mixed model framework. To assess the possible non-linearity of pollutant effects, dose-response-functions are also modelled as smooth functions in a second step.

The models for the AIRGENE data can be embedded into a more general framework of additive mixed models, where the additive components are modelled by P-Splines (penalized splines with a B-Spline basis, see Eilers and Marx, 1996), and the spline coefficients are estimated in a mixed model framework.

We use P-Splines rather than the truncated power basis usually used in this context, as in Ruppert, Wand and Carroll, 2003, or Ngo and Wand, 2004. This not only results in better numerical properties, but also allows us to independently choose the degree of the B-Splines and the order of the penalization, contrary to them being linked in the truncated powers approach. For the AIRGENE data we can thus use cubic B-Splines for smooth curves with 2nd order difference penalties, penalizing deviations from linearity as the natural default assumption.

We will use notation relating to our longitudinal data, but the extension to the more general additive mixed model case is obvious. A typical model for $y_{ij}$, the $j^{th}$ blood marker value of the $i^{th}$ patient, would be

$$y_{ij} = u_i + \sum_{l=1}^{w} x_{ijl} \beta_l + \sum_{k=1}^{t} f_k(s_{ijk}) + \varepsilon_{ij} \quad \text{with} \quad (1)$$

- $u = (u_1, \ldots, u_n) \sim N(\mathbf{0}, \sigma_u^2 I_n)$, the random person effects
- $x_{ij1}, \ldots, x_{ijw}$ values of the linear effect variables $x_1, \ldots, x_w$ for the $ij^{th}$ observation
- $\varepsilon = (\varepsilon_{11}, \ldots, \varepsilon_{1n_1}, \ldots, \varepsilon_{n_1}, \ldots, \varepsilon_{nn_n}) \sim N(\mathbf{0}, \sigma_\varepsilon^2 I_N)$, the error terms
- $f_1, \ldots, f_t$ smooth functions of continuous variables with

$$f_k(.) = \sum_{\nu=1}^{K(k)} \gamma_{k\nu} B_{\nu}^d(\cdot), \quad (2)$$
where the $B_{kl}^d(.)$ are $K(k)$ B-splines of degree $d$.

Omitting the $u_i$ for the moment, the penalized least squares problem for model (1) can then be written as

$$\min \| y - X\beta - \sum_{k=1}^t B_k \gamma_k \|_2^2 + \sum_{k=1}^t \lambda_k \gamma_k' D_{K(k)}^p \gamma_k$$

with (3)

- $D_h^p$ the $p^{th}$ order difference matrix of size $(h - p) \times h$:
  $$D_h^0 = I_h, D_h^p = \begin{bmatrix} -1 & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ -1 & 1 & \cdot & \cdot & \cdot \end{bmatrix}$$

- $\gamma_k = \Psi_k^{p,unp} \gamma_k^{p,unp} + \Psi_k^{p,pen} \gamma_k^{p,pen}$

We can now split the $\gamma$s into an unpenalized and a penalized part (see Fahrmeir, Kneib and Lang, 2004, for this idea for Bayesian P-Splines)

$$\gamma_k = \Psi_k^{p,unp} \gamma_k^{p,unp} + \Psi_k^{p,pen} \gamma_k^{p,pen}$$

The penalty terms then reduce nicely and our problem can be rewritten as

$$\min \| y - \tilde{X}\tilde{\beta} - \tilde{Z}\tilde{\gamma} \|_2^2 + \sum_{k=1}^t \lambda_k \gamma_k^{p,pen}' I_{K(k)} \gamma_k^{p,pen}$$

with fixed effects $\tilde{\beta}$ and random effects $\tilde{\gamma}$ in the mixed model

$$y = \tilde{X}\tilde{\beta} + \tilde{Z}\tilde{\gamma} + \varepsilon$$

with fixed effects $\tilde{\beta}$ and random effects $\tilde{\gamma}$ with a block diagonal covariance matrix with fixed variances $\sigma^2_{\gamma_k} = \sigma^2_{\gamma}/\lambda_k$. We can now easily re-include the other random effects in the model, appending $\tilde{\gamma}$ and $\tilde{Z}$ accordingly.
Centered plots for $f_k$ can be constructed setting up the design matrices for a grid of $s_k$, the mean values of the other continuous variables and the reference category of the categorical variables. After reparametrizing, resulting in the matrices $\tilde{X}_{f_k}$ and $\tilde{Z}_{f_k}$, we can estimate the vector of BLUPs for the grid as

$$\hat{f}_k = \tilde{X}_{f_k}\hat{\beta} + \tilde{Z}_{f_k}\hat{\gamma},$$

where $\hat{\beta}$ and $\hat{\gamma}$ are the estimated BLUPs for $\tilde{\beta}$ and $\tilde{\gamma}$.

Variability bands for the linear and smooth components can be computed using

$$C := \text{Cov}\left(\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} - \tilde{\gamma} \end{bmatrix}\right) = \sigma^2_{\epsilon} \begin{bmatrix} \tilde{X}'\tilde{X} & \tilde{X}'\tilde{Z} \\ \tilde{Z}'\tilde{X} & \tilde{Z}'\tilde{Z} + F \end{bmatrix}^{-1},$$

where $F = \text{blockdiag}(\sigma^2_{\epsilon}I_n, \sigma^2_{\epsilon}I_{K(1)}, \ldots, \sigma^2_{\epsilon}I_{K(t)})$ (see Ruppert, Wand and Carroll, 2003, for a derivation). An approximate 100(1−α)% confidence interval for the centered $f_k$ at a specific point $t_k$ can then for large numbers of observations be calculated as

$$\hat{f}_k(t_k) \pm z_{1-\frac{\alpha}{2}} \hat{\text{std}}\left(\hat{f}_k(t_k) - f_k(t_k)\right) = \hat{f}_k(t_k) \pm z_{1-\frac{\alpha}{2}} \sqrt{l_{t_k}^t\hat{C}l_{t_k}}$$

where $l_{t_k}$ is the corresponding row in $[\tilde{X}_{f_k} | \tilde{Z}_{f_k}]$ and $\hat{C}$ is constructed using the estimated variances. Partial residuals can be added to the plots as an additional diagnostic tool by summing the residuals and the componentwise fitted values.

4 Implementation

To our knowledge, this approach is not implemented in standard software as yet, although the implementation is described in Ngo and Wand, 2004, using a truncated lines basis. In our SAS macro, random intercepts, smooth, linear and categorical components can be named and the degree of the B-Splines, the order of the differences for the penalization and the number of knots can be chosen. Plots of the smooth components with variability bands and partial residuals as well as tests of the linear and categorical covariates as implemented in SAS proc mixed are available. The macro could potentially be used in a wide variety of contexts and can be obtained from the presenting author. The approach showed good results in the simulations we conducted.

5 Results

We decided on separate models for each blood marker and city, as climate and study period differ considerably. We built confounder models
TABLE 1. Effect estimates for log(IL-6) [pg/ml] in Stockholm.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>Std</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interception</td>
<td>0.1904</td>
<td>0.4304</td>
<td>0.6588</td>
</tr>
<tr>
<td>log(BNP) [pg/ml]</td>
<td>0.1336</td>
<td>0.0351</td>
<td>0.0001</td>
</tr>
<tr>
<td>BMI [kg/m^2]</td>
<td>0.0241</td>
<td>0.0085</td>
<td>0.0047</td>
</tr>
<tr>
<td>HDL [mg/dl]</td>
<td>-0.0064</td>
<td>0.0027</td>
<td>0.0188</td>
</tr>
<tr>
<td>Temperature [°C]</td>
<td>-0.0066</td>
<td>0.0051</td>
<td>0.1971</td>
</tr>
<tr>
<td>COPD/chronic bronchitis</td>
<td>no</td>
<td>-0.3045</td>
<td>0.1516</td>
</tr>
<tr>
<td>Reinfarction</td>
<td>no</td>
<td>-0.1557</td>
<td>0.1063</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>0</td>
<td>0.0177</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>0</td>
<td>0.1434</td>
</tr>
</tbody>
</table>

First, without air pollutants to allow for meaningful tests of pollutant effects, and then added one pollutant at a time to avoid collinearity, testing for a linear effect. Results were pooled subsequently using meta-analysis methodology.

The air pollution results of the AIRGENE study will be presented elsewhere. As an example for the analyses conducted, we here present the selected confounder model for one of the blood markers and one of the cities - for the log-transformed Interleukin-6 in Stockholm. This model was built in a forward step-wise procedure using the AIC to compare models. We selected chronic obstructive pulmonary disease / chronic bronchitis and reinfarction indicators (categorical), log(BNP) (a heart failure blood marker), body mass index (BMI), high density cholesterol (HDL) and average apparent temperature in the last 48 hours (linear), and time trend (smooth) as potential confounders. Results for the effect estimates of the linear and categorical predictors are shown in table 1. Higher values of BMI and BNP, lower values of HDL (the "good" cholesterol) and temperature as well as having COPD / chronic bronchitis or a reinfarction correspond to higher IL-6 values. Time trend was measured in days since start of the study, corresponding in Stockholm to September 03, 2003, to June 24, 2004. Figure 1 shows the estimated smooth time trend with 95% variability bands and partial residuals, estimated using cubic B-Splines with second order difference penalties.

6 Summary and Outlook

We have shown that additive mixed models with penalized splines estimated in the mixed model framework are extendable to P-splines using an additional reparametrization. This allows an independent choice of the B-Spline degree and the order of the penalty. We implemented this approach
in a SAS macro suitable for many applications and used it to analyze the longitudinal AIRGENE study.

We will next focus on testing of smooth components and pooling of smooth components across cities, which is motivated by estimation and testing of dose-response-functions for environmental factors in AIRGENE. A bootstrap test analogously to the truncated lines case in Coull, Schwartz, and Wand, 2001, will be implemented first. Afterwards we plan an extension of the exact likelihood ratio tests for penalized splines developed by Crainiceanu et al., 2005, using the truncated power basis, and a comparison of the two tests.

An extension of our approach to generalized models is also in progress, to allow models for additional diary data in AIRGENE where presence or absence of symptoms as well as health status in five categories was noted down daily. All extensions will be implemented and made available in SAS.

of Athens Medical School, Dep. of Hygiene a. Epi. (Greece): K. Katsouyanni (PI), A. Chalamandaris; 8. UHEL (Helsinki, Finland): M. Kulmala, P. Aalto, P. Paatero
Special thanks to Thomas Kneib and Susanne Breitner for fruitful discussions regarding additive mixed models.


Model selection for frailty structures

Il Do Ha \(^1\), Youngjo Lee \(^2\) and Gilbert MacKenzie\(^3\)

\(^1\) Daegu Haany University, South Korea
\(^2\) Seoul National University, South Korea
\(^3\) Centre of Biostatistics, University of Limerick, Ireland

Abstract: Frailty models are now widely used for analyzing multivariate survival data. An open question is how to select the most appropriate frailty structure supported by the data. Herein, we develop a procedure for selecting the optimal frailty structure from a set of (possibly) non-nested frailty models. Our focus is on the dispersion parameters which define the frailty structure. We propose two new AIC criteria: one based on the deviance for goodness of fit and the other on the extended restricted likelihood (ERL) of Lee and Nelder (1996). A simulation study shows that the AIC based on the extended restricted likelihood performs better.

Keywords: Model Selection; AIC; DIC; ERL; frailty models; \(h\)-likelihood.

1 Introduction

Various multivariate survival models based of different frailty structures have now been developed; for example, time dependent models (Yau & McGilchrist, 1998) and nested models (Sastry, 1997; Yau, 2001). Typically these are based on Cox’s (1972) proportional hazards (PH) model which has been generalized by including frailty components reflecting the complexity of the study design. Whilst all of these frailty components may be formally included in any model, not all may be supported by the data. Accordingly, it is important to develop a model selection procedure in which inference is focussed on the frailty variance components.

In this article the notion of focussing is intimately connected with nuisance parameter elimination, so that a focussed vehicle for inference, eg a likelihood, depends only on the parameters of interest - in this context - the variance components. Later we shall formalize this notion statistically in the context of \(h\)-likelihood inference.

Hierarchical generalized linear models (HGLMs), Lee & Nelder, (1996, 2001, 2005) extend, considerably, the class of random-effect models in the Exponential family. In a series of papers Lee & Nelder have developed inference for the parameters in these models based on their \(h\)-likelihood approach. In particular, for model selection, the deviance of goodness of fit, \(D\), in HGLMs can be used to form an information criterion, the Akaike
I.D. Ha et al. 209

information criterion (AIC, Akaike, 1973). Spiegelhalter et al. (2002) studied this type of deviance information criterion (DIC) in a Bayesian setting. Moreover, the extended restricted likelihood (ERL) of Lee and Nelder (1996, 2003) can also be used as a model selection criterion for dispersion parameters in the HGLM setting.

In this article we generalize the HGLM information criteria, D and ERL, to PH frailty models. Compared with classical random-effect models such as HGLMs, inference in semi-parametric frailty models is complicated by censoring and by the presence of a nonparametric baseline hazard function.

2 PH Frailty Models

Let \( T_{ij} (i = 1, \ldots, q, \ j = 1, \ldots, n_i, \ n = \sum_i n_i) \) be the survival time for \( j \)th observation on the \( i \)th subject and \( C_{ij} \) be the corresponding censoring time. Then the observable data become \( s_{ij} = \min(T_{ij}, C_{ij}) \) and \( \delta_{ij} = I(T_{ij} \leq C_{ij}) \), where \( I(\cdot) \) is the indicator function. Denote by \( U_i \) the unobserved frailty random variable for the \( i \)th subject. Let \( V_i = \log U_i \) and \( v_i = \log u_i \). A frailty model is described as follows. Given \( V_i = v_i \), the conditional hazard function of \( T_{ij} \) is of the form:

\[
\lambda_{ij}(t_{ij}|v_i) = \lambda_0(t_{ij}) \exp(x_{ij}^T \beta + v_i) \tag{1}
\]

where \( \lambda_0(\cdot) \) is an unspecified baseline hazard and \( \beta \) is a \( p \times 1 \) vector of fixed effects associated with fixed covariates \( x_{ij} = (x_{ij1}, \ldots, x_{ijp})^T \). The log-frailties, \( V_i i = 1, \ldots, q \), are assumed to be independent and identically distributed random variables having a frailty parameter \( \alpha \). Even though the results of this paper can be applied to non-normal frailties, for simplicity of argument we employ the normal distribution for \( V_i \), which is useful for modelling correlated frailties.

To illustrate, we consider one-component model \( X\beta + Zv \), which can be easily extended to a multi-component model as follows:

\[
X\beta + Z_1v^{(1)} + Z_2v^{(2)} + \cdots + Z_kv^{(k)} \tag{2}
\]

\( X \) is the \( n \times p \) model matrix, \( Z_r (r = 1, 2, \ldots, k) \) are the \( n \times q_r \) model matrices corresponding to the \( q_r \times 1 \) frailties \( v^{(r)} \), and \( v^{(r)} \) and \( v^{(l)} \) are independent for \( r \neq l \). Let \( Z = (Z_1, Z_2, \ldots, Z_k) \), \( v = (v^{(1)}T, v^{(2)}T, \ldots, v^{(k)}T)^T \), \( \alpha = (\alpha_1, \ldots, \alpha_k)^T \), and \( q = \sum_r q_r \). We use \( \alpha \) and \( \rho \) to represent dispersion parameters in the frailty distribution. Then the multi-component model can be substituted directly into (1) leading to a straightforward extension.

3 \( h \)-likelihood Inference & ERL

Let \( \ell = \ell(\beta, \theta) \) be a likelihood, either an \( h \)-likelihood, \( h \), or a marginal likelihood, \( m = \log \left\{ \int \exp(h) \ dv \right\} \), with nuisance parameters \( \theta \). Lee and
Nelder (2001) considered a function $p_\theta(\ell)$, defined by
\[ p_\theta(\ell) = \left[ \ell - \frac{1}{2} \log \det \left\{ A(\ell, \theta)/(2\pi) \right\} \right]_{\theta = \hat{\theta}} \]

where $A(\ell, \theta) = -\partial^2 \ell / \partial \theta^2$ and $\hat{\theta}$ solves $\partial \ell / \partial \theta = 0$. The function $p_\theta(\cdot)$ produces an adjusted profile likelihood, eliminating nuisance effects $\theta$, which can be fixed effects $\beta$ or random effects $v$ or both.

In general, $p_\beta(m) \simeq r$, the restricted likelihood (REML) to the first order (Cox & Reid, 1987), $p_v(h)$ is the first-order Laplace approximation to $m$ (i.e. $p_v(h) \simeq m$) and $p_{\beta,v}(h) \simeq p_\beta(m)$ (Lee & Nelder, 2001). In principle, one should use the h-likelihood, $h$, for inferences about $v$; the marginal-likelihood, $m$, for $\beta$; and the restricted likelihood, $p_\beta(m)$, for the dispersion parameters. When $m$ is numerically difficult to obtain, we may use $p_v(h)$ and $p_{\beta,v}(h)$ as approximations to $m$ and $p_\beta(m)$, respectively.

Lee & Nelder (2003) called $p_{\beta,v}(h) \simeq r$ the extended restricted likelihood (ERL) - the extension being to the (REML-based) elimination of the random effects, $v$.

## 4 Deviances, Extensions & Information Criteria

Nelder and Lee (1996) suggested the deviance of goodness of fit, $D$, as a model selection criteria, namely
\[ D = D(y; \hat{\mu}) = -2\{ \ell_1(\hat{\mu}; y|v) - \ell_1(y; y|v) \} \]

in the HGLM class of random effect models. It may be shown (Ha et al, 2005) that in general PH survival models with parametric frailty give rise to the same likelihood as Poisson HGLMs, so that (4) remains a vehicle for model selection in the frailty setting.

The deviance $T_d$, based upon the ERL, is given by
\[ T_d = -2\{ p_{\beta,v}(h) \} \]

is also a candidate model selection criterion (Lee & Nelder, 2001) and, notably, it is focussed solely on the variance components, the fixed effects and random effects having been eliminated by the REML adjustment.

As they stand these are not information criteria since we have not made any formal adjustment for nuisance parameter elimination. Moreover, we have not yet discussed the elimination of the nuisance function $\lambda_0(t)$. Nevertheless, the idea of comparing the performance of these unfocussed and focussed model selection criteria begins to intrude.

It may be shown that the two deviances can be extended to PH survival frailty models via the following information criteria:
\[ \text{AIC}(D^*) = D + 2p_D \]
where the estimated degrees of freedom, \( d.f. = N - p_D \), \( N = \sum_k \sum_{ij} I\{ (i,j) \in R(y_{ik}) \} \) is the number of observations \( y_{ij,k} \) for the equivalent Poisson HGLM, and \( p_D = \text{trace}(H^{-1}H^*) \) where: \( H = A(h, \theta) \), \( H^* = A(\ell_1, \theta) \) and \( \theta = (\lambda_0^T, \beta^T, v^T)^T \).

Similarly, the focussed model selection criterion for the dispersion parameters can be extended first to a new deviance:

\[
T^*_d = -2\{p_{\beta,v}(h^*)\}
\]

and then to the focussed information criterion:

\[
\text{AIC}(T^*_d) = T^*_d + 2p_T
\]

Notice that \( T^*_d \) is the ERL based on \( p_{\beta,v}(h^*) \) which eliminates \( \beta, v \) from \( h^* \), the profile \( h \)-likelihood from which the nuisance function \( \lambda_0(t) \) has already been eliminated - a strategy proposed by Ha, Lee & Song (2001) and Ha & Lee (2005). Since \( T^*_d \) is by construction the ERL for the dispersion parameters, \( p_T \) turns out to be the number of dispersion parameters or variance components in the model.

5 Results

5.1 Mammary Tumour Data

We re-analyze the data of Gail et al. (1980) on multiple occurrences of mammary tumours for 48 female rats. The observations are the times to the development of a mammary tumour for 23 female rats in the treatment group and 25 female rats in the control group. Initially, 76 rats were injected with a carcinogen, and each rat was treated with retinyl acetate for the next 60 days. Some 48 rats were tumour-free after 60 days. These rats were randomly assigned to continued retinoid prophylaxis or to the control group, where they received no treatment. The main objective of the study was to evaluate treatment.

The survival time from the initial carcinogen injection, \( T_{ij} (j = 1, \ldots, n_i) \) is then calculated as \( t_{i,j} - t_{i,j-1} \), where \( t_{i,j} \) with \( t_{i,0} = 0 \) is the \( j \)th tumour occurrence time of the \( i \)th rat, i.e., the gap time between tumour recurrences. Survival times on the same rat may be correlated due to shared genetic or environmental effects and this correlation can be modelled by a shared frailty. Moreover, the frailty of each rat may not be constant, but may vary stochastically over the gap times. Following Yau and McGilchrist (1998), we consider AR(1) frailty models for such dependency. Here we model a single fixed covariate \( x_{ij} (= 1 \text{ for treatment and} = 0 \text{ for control}) \) in the following five models, \( \lambda_{ij}(t_{ij}|v) = \lambda_0(t_{ij}) \exp(\eta_{ij}) \) where \( \eta_{ij} \) allows for the covariate and/or the frailty structures in models below:
TABLE 1. Deviance results for the mammary tumor data.

<table>
<thead>
<tr>
<th>Model</th>
<th>$T_d^*$</th>
<th>$p_T$</th>
<th>AIC($T_d^*$)</th>
<th>$D$</th>
<th>$p_D$</th>
<th>AIC($D$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1 (Cox)</td>
<td>1946.8</td>
<td>0</td>
<td>7.0</td>
<td>1204.8</td>
<td>58</td>
<td>28.0</td>
</tr>
<tr>
<td>M2 (R)</td>
<td>1939.1</td>
<td>1</td>
<td>1.3</td>
<td>1147.4</td>
<td>78.8</td>
<td>12.2</td>
</tr>
<tr>
<td>M3 (AR(1)*)</td>
<td>1946.7</td>
<td>1</td>
<td>8.9</td>
<td>1204.8</td>
<td>58.0</td>
<td>28.0</td>
</tr>
<tr>
<td>M4 (AR(1))</td>
<td>1935.8</td>
<td>2</td>
<td>0</td>
<td>1080.4</td>
<td>106.2</td>
<td>0</td>
</tr>
<tr>
<td>M5 (R+AR(1))</td>
<td>1935.8</td>
<td>3</td>
<td>2.0</td>
<td>1080.4</td>
<td>106.2</td>
<td>0.0</td>
</tr>
</tbody>
</table>

R, rat frailty; AR(1), AR(1) frailty; AR(1)*, AR(1) frailty with $\rho = 0$; $T_d^* = -2[p_{\beta,v}(h^*)]$; $p_T$, the number of frailty parameters; $D$, deviance; $p_D = \text{trace}(H^{-1}H^*)$; AIC, Akaike information criterion differences where the smallest AIC is adjusted to be zero.

M1 (Cox): $\eta_{ij} = x_{ij}\beta$,
M2 (R): $\eta_{ij} = x_{ij}\beta + v_i$ with $v_i \sim N(0, \alpha_1)$,
M3 (AR(1)*): $\eta_{ij} = x_{ij}\beta + e_{ij}$ with $e_{ij} \sim N(0, \alpha_2)$,
M4 (AR(1)): $\eta_{ij} = x_{ij}\beta + v_{ij}$ with $v_{ij} \sim \text{AR}(1)$,
M5 (R+AR(1)): $\eta_{ij} = x_{ij}\beta + v_i + v_{ij}$ with $v_i \sim N(0, \alpha_1)$ and $v_{ij} \sim \text{AR}(1)$.

Here $v_{ij} \sim \text{AR}(1)$ means that $v_{ij} = \rho v_{i,j-1} + e_{ij}$, $e_{ij} \sim N(0, \alpha_2)$ and $|\rho| < 1$.

In this paper we select the model which has the smallest AIC value among these models. For ease of comparison and ranking of candidate models, we have set the smallest value to be zero. In Table 1 we report the AIC differences, not the AIC values themselves.

Overall, the results in Table 1 suggest that $p_D$ may not reflect model complexity properly when the variance of the frailties is near zero. Thus, for model selection related to ($\alpha, \rho$) we should prefer AIC($T_d^*$). If the AIC difference is larger than 1~2 it is considered to be significant, and if the difference is less than 1 it is not. Using this criterion, AIC($T_d^*$) selects M4 as the final model, while the AIC($D$) cannot select between M4 and M5.
TABLE 2. Simulation Results: percentage correct selection.

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Criterion</th>
<th>True model</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SM1</td>
<td>SM2</td>
</tr>
<tr>
<td>Case 1: (q, n_i) = (80, 5)</td>
<td>AIC(T^*_f)</td>
<td>94</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>AIC(D)</td>
<td>10</td>
<td>99</td>
</tr>
<tr>
<td>Case 2: (q, n_i) = (20, 20)</td>
<td>AIC(T^*_f)</td>
<td>93</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>AIC(D)</td>
<td>46</td>
<td>100</td>
</tr>
</tbody>
</table>

5.2 Simulation Study

Here we report the results of a small simulation study based on 100 replications to evaluate the performance of the two AICs proposed above. We consider two non-nested models:

SM1: \( \eta_{ij} = x_{ij} \beta + v_i \) with \( v_i \sim N(0, \alpha_1) \),
SM2: \( \eta_{ij} = x_{ij} \beta + v_{ij} \) with \( v_{ij} \sim \text{AR}(1) \).

In order to generate the data from SM1 and SM2, we used exponential distributions for both survival and censoring times, with a censoring rate of 20%. Here we set \( \lambda_0(t) = 1.0 \) and \( \beta = -1.0 \), \( x_{ij} \) to 0 for the first \( q/2 \) subjects, to form the control group, and \( x_{ij} \) to 1 for the remaining \( q/2 \), to form the treatment group. We also set \( \alpha_1 = \alpha_2 = 0.5 \) and \( \rho = 0.7 \). We anticipated that the distinction between SM1 and SM2 might be difficult to detect with small \( n_i \). Accordingly, we used a sample size of \( n = \sum_{i=1}^{q} n_i = 400 \) in two scenarios \( (q, n_i) = (80, 5) \) and \( (q, n_i) = (20, 20) \). From the 100 replications we computed the two AICs. Table 2 shows the rate of selection of the true model among 100 replications. The results clearly favour the focussed information criterion.

6 Final Remarks

Two new information criteria for model selection in non-nested frailty models have been defined. One criterion is based on a general deviance approach while the other is based on a focussed deviance approach; where the focus is on the variance components which define the frailty model structure. The latter relies on ERL inference in which REML-like arguments are used eliminate the nuisance parameters - fixed effects, random effects and arbitrary baseline hazard components - from the \( h \)-likelihood. The use of the criteria have been illustrated in the analysis of a well known multivariate data set.
Selection for frailty models

and by means of a small simulation study. Overall the results favour the use of the focussed criterion based on ERL inference.

It will be recognized that we are only in the early stages of developing ERL-based information criteria for frailty model selection. The idea of nuisance parameter elimination is, of course, not new; but the novelty of this paper lies in its sustained use of ERL arguments to produce, for the first time, a focussed information criterion applicable to a popular class of frailty models.

References


An empirical model and scenario analysis of nitrous oxide emissions from a fertilised and grazed temperate grassland site

Michael J. Hawkins¹, Bernard P. Hyde², Michael Ryan², R.P.O. Schulte² and John Connolly¹

¹ UCD School of Mathematical Sciences, Belfield, Dublin 4, Ireland.
² Teagasc, Johnstown Castle Research Centre, Wexford, Ireland.

Abstract: This paper describes an empirical model of soil-evolved nitrous oxide emissions inferred from data gathered in a two-year rotational grazing study of grassland emissions. The model was used to simulate daily and annual emissions for the nine-year period between 1994 and 2002 under a variety of possible fertiliser application scenarios. Results reveal a high interannual variability of emissions, increasing with the amount of nitrogen fertiliser applied. The results of bootstrapping tests to statistically compare the simulated emissions from the application scenarios are also presented.

Keywords: Nitrous oxide; soil; generalized linear mixed model; scenario analysis; bootstrap.

1 Introduction

Nitrous oxide (N₂O) is an atmospheric gas that contributes to global warming, acid rain and the depletion of stratospheric ozone. It is generally accepted that soils are the most important source of N₂O globally (Bouwman et al., 1993). However, high spatial and temporal variability of soil-evolved N₂O emissions (due in large part to their dependence on climatic and weather conditions) makes the task of quantifying and studying emissions at the field scale over the long- and medium-term by field measurements prohibitively intensive. Models therefore provide a useful way of studying field-scale emissions.

Most published soil-emissions models can be categorised as mechanistic or empirical. In general, the former require more input data than the latter. The motivation of the present work was to derive a model of soil-evolved emissions that required few inputs. The model was then to be used to investigate the behaviour of annual emissions. We estimate a GLMM with serial correlation from a longitudinal data set with nine subjects and 133 subject measurements (McCulloch & Searle, 2001). The fitted model and
historic weather series are then combined to investigate annual emissions under a variety of fertiliser regimes and climatic conditions.

2 Materials and Methods

2.1 Gas sampling and analyses

The data on which the empirical model is based were collected in a two-year rotational grazing experiment that was carried out in Johnstown Castle, Wexford, Ireland, between November 2001 and November 2003. The experiment is described in full in a paper currently in review (Hyde et al, in review). The experimental field was divided into nine plots and three fertiliser regimes were imposed, each on three replicate plots: 0, 225 and 390 kg N/ha/year. Rotational grazing (periodically moving livestock to fresh paddocks, to allow pastures to regrow) was practised using three separate herds of steers.

Animals grazed for one week in the first replicate of each treatment and fertiliser was applied to these plots following removal of the animals. The herds were then moved to graze for one week on the second replicate plot of each treatment. Fertiliser was applied to these plots following removal of the animals, and the process continued cyclically in this manner (a 21-day grazing rotation) for the duration of the grazing season. Nitrous oxide measurements on all nine plots began on the same day that animals were removed from the plots grazed in that rotation. Nitrogen fertiliser was then applied to all plots and emission measurements continued daily for a period of five days after application after which the intensity of measurement was reduced to two or three times per week and subsequently to once per week. Additional measurements were made following heavy rainfall events.

2.2 Statistical model

A model was fitted that allows one to estimate field-scale emissions based on soil temperature, volumetric soil moisture, the amount of fertiliser nitrogen (N) applied, and the time elapsed since the fertiliser application. This subset of the set of variables that influence emissions includes the variables that are understood to be of primary importance to the emissions process (Schmid et al, 2001). The modelled response variable for each plot was the arithmetic mean of eight daily static-chamber measurements of \( \mathrm{N}_2\mathrm{O} \), scaled to units of grams/ha/day. Let \( y_{ij} \) denote the response from the \( i \)th plot on the \( j \)th measurement occasion, where \( i = 1 \ldots 9 \), and \( j = 1 \ldots 133 \). The model equation is:

\[
E(y_{ij}|u_i) = \exp(x_{ij}^T \beta + u_i) = \mu_{ij}
\]

Where \( E(\cdot) \) denotes expectation; \( \exp(\cdot) \) is the exponential function; \( x_{ij}^T = \begin{bmatrix} 1 & x_{1ij} & x_{2ij} & \cdots & x_{pij} \end{bmatrix} \) is a vector of independent variables; \( u_i \) is a \( NID(0, \sigma_u^2) \)
random effect of the $i$th plot; and $\beta_{(p+1)\times1}$ is vector of fixed effects regression parameters, including an intercept. Heteroscedasticity in the data is accommodated with a quadratic variance function, $\mu_{ij}^2$, and a constant coefficient of variation $\phi$, so the modelled variance of emissions, conditional on plot effects, is:

$$\text{Var}(y_{ij}|u_i) = \phi \mu_{ij}^2$$  \hspace{1cm} (2)

We model serial correlation induced by the repeated-measures in the experimental design as continuous-time first-order autoregression (Jones, 1993). This structure accommodates unequally-spaced repeated measurements. Since we assume independence across plots,

$$\text{Cov}[y_{mi}|u_m, y_{nj}|u_n] = \phi \mu_{mi} \mu_{nj} \rho_{d_{ij}}$$

for $m = n$ and 0 otherwise. Here $\rho_{d_{ij}}$ represents the correlation between observations separated by a time interval of $d_{ij}$, where:

$$d_{ij} = |c_{mi} - c_{mj}|$$

and the vector $c$ indexes each observation in the response vector by the number of days since the start of the experiment.

2.3 Scenario analysis

Using the fitted prediction equation and observed and estimated time series of daily climatic data, we compiled estimates of daily N$_2$O emissions (g/ha/day), for the period January 1, 1994, to December 31, 2002, inclusive, for three annual rates of fertiliser application: 0, 170 and 350 kg N/ha/yr. Many different simulations are possible with the model but the following scenarios illustrate an interaction between soil moisture status and fertiliser N application date:

MONTHLY: Fertiliser is applied on the first day of each month of an application season, February – September, inclusive.

DRIEST: Fertiliser is applied on the day in the first week of the application months (February – September, inclusive) when soil moisture content is at a weekly minimum.

WETTEST: Fertiliser is applied on the day in the first week of the application months (February – September, inclusive) when soil moisture content is at a weekly maximum.

The historic daily soil moisture series used in the scenario simulation was estimated using a model for predicting soil moisture deficit (Schulte et al, 2005). The soil temperature series was obtained from Met Éireann.
FIGURE 1. Estimated differences in annual emissions (solid circle) and their associated approximate 95% confidence limits (circle) for rates of 170 kg N/ha/year (upper graph) and 350 kg N/ha/year (lower graph) for various scenario comparisons. Note: D = Driest, W = Wettest and M = Monthly. “W - D” reads “Average annual emissions from the Wettest scenario minus average annual emissions from the Driest scenario”.

-0.5 0.0 0.5 1.0 1.5 2.0 2.5
W - M D - M W - D

-0.2 0.0 0.2 0.4 0.6 0.8 1.0
W - M D - M W - D
2.4 Comparing scenarios

The output from the scenario analysis sheds some light on the possible behaviour of annual emissions for the period 1994–2002 (table 1). However, it is also of interest to test hypotheses using the model and the climatic data series. We use the bias-corrected and accelerated bootstrap for this purpose (Efron & Tibshirani, 1993). The effect of an application of fertiliser N on annual emissions changes from year to year because of the interaction between fertiliser and climatic factors. We therefore compare scenarios using model estimates of annual emissions averaged over the period of the simulation (1994-2002) inclusive. Results of the bootstrapping are shown in figure 1.

3 Results and conclusions

The interannual variability of annual emissions due to climatic influence increases with fertiliser rate for all scenarios investigated here (table 1). In general, the estimated differences between scenarios also increase with fertiliser rate (fig. 1).

<table>
<thead>
<tr>
<th>Scenario</th>
<th>N applied (kg/ha/yr)</th>
<th>Mean Emission (kg N2O-N/ha/yr)</th>
<th>CV(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MONTHLY</td>
<td>0</td>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>170</td>
<td>16.2</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>350</td>
<td>18.7</td>
<td>23</td>
</tr>
<tr>
<td>DRIEST</td>
<td>0</td>
<td>6</td>
<td>6.3</td>
</tr>
<tr>
<td></td>
<td>170</td>
<td>16</td>
<td>14.5</td>
</tr>
<tr>
<td></td>
<td>350</td>
<td>18.3</td>
<td>22</td>
</tr>
<tr>
<td>WETTEST</td>
<td>0</td>
<td>6.1</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>170</td>
<td>16.7</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>350</td>
<td>19.9</td>
<td>26.4</td>
</tr>
</tbody>
</table>

Mitigation strategies based on the timing of fertiliser application (which attempt to exploit the interaction between applied fertiliser and climatic conditions) can be investigated with the fitted model. The average annual emission from the Monthly scenario exceeded that of the Driest scenario. This reduction was of the same order of magnitude for both fertiliser rates simulated here, and was not statistically significant at the 5% significance level at the 170 kg/ha/year fertiliser rate (fig. 1). However, the average annual emission from the Wettest scenario exceeded that of the Monthly scenario at both rates of applied fertiliser by a statistically significant amount at the 5% significance level. Similarly, emissions from the Wettest scenario
were significantly higher than those from the Driest scenarios for both rates of applied fertiliser. The fitting process and type of model described in this paper has not been previously used in modelling emissions. The model described here has advantages over other published models in that it requires a small number of easily obtainable site-specific parameters to predict emissions. Future work will extend the study to examine emissions from different soil types.

Acknowledgments: Special Thanks to the EPA, Ireland.

References


Exponential family mixtures in total quality health care monitoring

Karl Heiner$^1$ and John Hinde$^2$

$^1$ State University of New York at New Paltz, USA
$^2$ National University of Ireland, Galway, Ireland

Abstract: In Health Care Performance Monitoring, one approach requires health care providers to report values for variables that are designed to measure quality of care at their health care facility. The purpose of this activity is to economically gather relevant data and to engage the provider in the Total Quality Management (TQM) approach, in which the people involved directly in the health care delivery actively participate in the assessment of their activities. Because the results of the performance monitoring are frequently made public, potentially being used by those seeking care and in the news media to bring attention to exemplary care providers and particularly low ranked providers, independent Peer Review Organizations (PROs) that originally had acted as a gold standard, audit the self-evaluations. More recently, health care quality experts have hypothesized that the self-evaluations may be even more reliable than the ones performed by the PROs. We provide an example where logistic regression mixture models are used to classify providers based on level of agreement between self-assessment and independent peer review assessment. This analysis identifies providers that bring more information to bear on quality assessment than that available to peer reviewers reading medical records and to providers needing training in the use of the quality monitoring protocols.

Keywords: generalized linear mixed models; healthcare quality monitoring; exponential family mixtures

1 Background

In New York State (NYS), the Department of Health’s AIDS Institute is charged with monitoring the quality of care received by patients with HIV. More than one fifth of the reported AIDS cases in the United States reside in NYS and receive care from hospitals, community health centers and drug treatment centers. The AIDS Institute has been monitoring the care provided to HIV patients since 1992 and more recently has published reports comparing the performance of the approximately 130 health care providers. The monitoring is based on a number of indicators developed by an advisory committee composed of physicians representing a large proportion of the hospitals providing care. This practice of involving those being assessed is consistent with one of the main recommendations of the Royal
Statistical Society’s Working Party on Performance Monitoring, as is the
development of detailed protocols for reviewers to follow. For more than a
decade, the reviewers have been nurses supervised by a head nurse and a
physician. It is notable that these performance indicators are not measures
of outcomes, but rather measures of process determined by review of the pa-
tient medical records. Some key quality indicators focus on monitoring and
response to changes in viral load and cd4 levels (ARV management), pneu-
monia (PCP prophylaxis), tuberculosis (PPD), and gynecological exams
for women (Pelvic). The protocols are rather involved, which introduces
the opportunity for measurement error.

For example, in the ARV management protocol, the evaluator must first
determine if a patient is stable or unstable, based on the results of a se-
ries of recent lab tests, primarily cd4 counts and viral load assays. If the
patient is stable, timely testing and evaluation of these lab results is suf-
ficient to document appropriate ARV management. If the patient is not
stable, appropriate management depends on more frequent patient visits,
prescribed treatments and the evaluation of the patient’s response to treat-
ment. Clearly, measurement error possibilities are numerous.

Annually, stratified random samples are drawn from patient lists submitted
by facilities to the AIDS Institute. The samples are stratified by gender
to insure a certain level of precision for the indicators that apply only
to women. Historically, facilities were visited by a team of nurses from
an independent Peer Review Organization (PRO) contracted to apply the
protocols developed by the advisory committee to the records of the selected
patients. For about half of the facilities, this is still the practice.

More recently, a number of the facilities have elected to participate in a
program of self-assessment, designed to yield results on the same indica-
tors as those monitored by the PRO. The purpose of this activity is to
economically gather relevant data and to engage the provider in the Total
Quality Management (TQM) approach, an approach used with success in
industrial quality control. Here, the people involved directly in the health
care delivery actively participate in the assessment of their activities. Be-
cause the results of the performance monitoring are to be made public,
potentially being used by those seeking care and in the news media to
bring attention to exemplary care providers and particularly low ranked
providers, independent Peer Review Organizations (PROs) audit the self-
evaluations. These PROs were originally thought of as a gold standard, but
more recently, health care quality experts have hypothesized that the self-
evaluations may be more reliable than the ones performed by the PROs,
in as much as the health care provider is more familiar with a patient’s
condition and history than an independent reviewer whose information is
constrained to the written record of care. In this case, latent variable mod-
ing is useful. Here the latent variable would be the true assessment for
each case measured with error by both the provider and the peer.

One may fit a number of generalized linear mixed models to the resulting
data in an effort to shed light on issues such as self assessment bias, agreement between assessments conducted by the providers and the PRO, and the possibility of using latent variables influenced by both sources of data. As an example, we fit a series of logistic regression mixture models to group providers by degree of agreement with independent peer reviewers. Based on these models, the NYS AIDS Institute intervenes hoping to increase agreement in subsequent review periods.

2 Data

The measures of interest to the facilities, the Department of Health, and the public (the response variables) are either dichotomous or measures of time between events. Facility covariates include facility type and location. Patient descriptors include gender, age, race, and risk group. Data are available for each facility for a varying number of years of participation, and in more recent years, many facilities have both their self-assessment results and results from the assessment performed on a sub-sample of patients by the PRO. In our example, we have summarized data by provider and are interested in level of agreement between self-assessment and peer assessment. Below we will model this agreement, a Bernoulli random variable where success is when both raters agree and a failure when they do not. It would also be possible to model the four possible categories of agreement (YY, NN, YN, and NY).

3 A Model for Agreement Classification

Our example is a non-parametric random effects mixture model.

\[ Y \mid Z \sim \text{binomial}(m, \pi) \]

\[ \text{logit}(\pi) = Z + X\beta \]

where \( X \) is a matrix of predictors, potential categorical or continuous and \( \beta \) is a vector of coefficients. \( Z \) is a random effect that we have taken to be a finite discrete distribution with \( K \) mass points \( Z_k \) with masses \( p_k \). This \( K \) component mixture model is fitted with code in the npmlreg R package (Einbeck and Hinde, 2005, Einbeck et al., 2006) that uses the EM algorithm and Gaussian quadrature.

4 Results

In our simple example, we fit models with only an intercept, varying the number of groups, \( K \), from one to six. Disparity, \(-2 \log \text{likelihood}\), for each of these six models is 298.4, 249.7, 247.6, 247.6, 246.2 and 246.2, respectively. Mixture proportions associated with the \( K \) groups in each model
are \{\{1.00\}, \{0.30, 0.70\}, \{0.27, 0.57, 0.16\}, \{0.06, 0.21, 0.059, 0.14\}, \ldots \}. We are inclined toward the three-group solution. A table of the sixty-four providers with probabilities of group membership and predicted agreement proportion was ordered by predicted agreement proportion. Group 1 are those providers where the provider and the peer agreement is nearly perfect. In group 2 agreement is high, while group 3 consists of providers where either the peer or the provider require additional guidance in the use of the measurement protocol. For illustrative purposes, we took a one in eight systematic sample to produce Table 1.

Managers at the NYS Health department have found the classification produced by this kind of model and the predicted agreement proportions helpful when consulting with the individual providers and the peer review organization on the use of the quality of care assessment protocols.

**References**


---

<table>
<thead>
<tr>
<th>Provider</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>pred.agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>59</td>
<td>0.000</td>
<td>0.276</td>
<td>0.724</td>
<td>0.99</td>
</tr>
<tr>
<td>8</td>
<td>0.003</td>
<td>0.478</td>
<td>0.519</td>
<td>0.99</td>
</tr>
<tr>
<td>14</td>
<td>0.027</td>
<td>0.612</td>
<td>0.361</td>
<td>0.98</td>
</tr>
<tr>
<td>10</td>
<td>0.017</td>
<td>0.945</td>
<td>0.038</td>
<td>0.94</td>
</tr>
<tr>
<td>53</td>
<td>0.043</td>
<td>0.957</td>
<td>0.000</td>
<td>0.93</td>
</tr>
<tr>
<td>62</td>
<td>0.162</td>
<td>0.822</td>
<td>0.016</td>
<td>0.92</td>
</tr>
<tr>
<td>57</td>
<td>0.521</td>
<td>0.473</td>
<td>0.005</td>
<td>0.86</td>
</tr>
<tr>
<td>40</td>
<td>0.925</td>
<td>0.075</td>
<td>0.000</td>
<td>0.77</td>
</tr>
</tbody>
</table>
The zero-adjusted Inverse Gaussian distribution as a model for insurance claims

Gillian Heller\textsuperscript{1}, Mikis Stasinopoulos\textsuperscript{2} and Bob Rigby\textsuperscript{2}

\textsuperscript{1} Dept of Statistics, Macquarie University, Sydney, Australia, gheller@efs.mq.edu.au
\textsuperscript{2} STORM, London Metropolitan University, d.stasinopoulos@londonmet.ac.uk and r.rigby@londonmet.ac.uk

Abstract: We introduce a method for modelling insurance claim sizes, including zero claims. A mixed discrete-continuous model, with a probability mass at zero and an Inverse Gaussian continuous component, is used. The Inverse Gaussian distribution accommodates the extreme right skewness of the claim distribution. The model explicitly specifies a logit-linear model for the occurrence of a claim; and log-linear models for the mean claim size (given a claim has occurred); and the dispersion of claim sizes (given a claim has occurred). The method is illustrated on an Australian motor vehicle insurance data set.

Keywords: Inverse Gaussian model; zero-adjusted; insurance claims; gamlss.

1 Introduction

The purpose of modelling claim sizes on insurance policies is to price premiums accurately, and to estimate the risk of extreme claim events. In a fixed period, a policy will either experience a claim, which is a non-negative amount typically having an extremely right-skewed distribution, or no claim, in which the claim amount is identically zero. The distribution of the claim size is then mixed discrete-continuous: a continuous, right-skewed distribution mixed with a single probability mass at zero. In this respect the phenomenon is similar to rainfall, which is either identically zero on a dry day, or a continuous non-negative size on a wet day.

1.1 Models for insurance claims

Much attention has been paid in the actuarial literature to alternative distributions for claim sizes (e.g. Hogg and Klugman (1984)) and some authors have developed regression models (usually generalized linear models) for explaining claim sizes as a function of risk factors (e.g. Haberman and Renshaw (1996)). All of these are models for claim sizes in the subclass of policies which had a claim in the period of observation.
Jørgensen and de Souza (1994) and Smyth and Jørgensen (2002) considered models for claim sizes, including the zero claims. These are based on the Tweedie distribution, which may be characterised as a Poisson sum of Gamma random variates. A problem with the Tweedie distribution model is that the probabilities at zero can not modeled explicitly as a function of explanatory variables; and as we shall see in the example, the Gamma distribution is inadequate for modelling the extreme right-skewness which is present in our data.

2 The zero-adjusted Inverse Gaussian model

Let \( y_i \) = size of claim on \( i^{th} \) policy, \( i = 1, \ldots, n \). We can write the distribution of \( y \) as a mixed discrete-continuous probability function:

\[
    f(y) = \begin{cases} 
        1 - \pi & y = 0 \\
        \pi \cdot g(y) & y > 0 
    \end{cases}
\]

where \( g(y) \) is the density of a continuous, right-skewed distribution and \( \pi \) is the probability of a claim.

2.1 Continuous part of the model

The extreme right skewness of claims distributions has been well documented. Candidate distributions within the exponential family are the Gamma and Inverse Gaussian distributions.

Motor vehicle insurance example We illustrate the method on a class of motor vehicle insurance policies from an Australian insurance company in 2004-05. There were 67,856 policies, of which 4,624 (6.8%) had at least one claim in the period of observation. Of these, 4,333 policies (6.4%) had one claim, and the remaining 291 policies (0.4%) between 2 and 4 claims. The maximum claim size was $56,000. A histogram of the non-zero claims, and the pdfs of the fitted Gamma and Inverse Gaussian distributions are shown in Figure 1. (For clarity of display the horizontal axis has been truncated, at $15,000. Sixty-five observations were omitted.) The Gamma clearly does not reproduce the shape of the observed claim size distribution; the Inverse Gaussian looks to be a far better fit, accommodating both the mode near zero and the extremely long tail of the distribution.

The density of the inverse Gaussian is:

\[
    g(y) = \frac{1}{\sqrt{2\pi y^3\sigma}} \exp\left[-\frac{1}{2y}\left(\frac{y - \mu}{\mu\sigma}\right)^2\right] \quad y > 0
\]

which has \( E(y) = \mu \) and \( Var(y) = \sigma^2\mu^3 \). The use of the Inverse Gaussian distribution for modelling claim sizes has been recommended by, for example, Berg (1994).
2.2 Discrete part of the model

The obvious model for the probability of a claim is the Bernoulli. Let \( w_i \) be a binary variable indicating the occurrence of at least one claim, and \( \pi_i \) be the probability of at least one claim, on policy \( i \). Note that the occurrence of more than one claim in the period of observation is rare. Then

\[
f(w_i) = \pi_i^{w_i} (1 - \pi_i)^{1 - w_i}
\]

However, we have to correct for the typical feature of policy-level data, that not all policies have been in force for the entire period of observation.

Let \( t_i = \) exposure of policy \( i \), \( 0 < t_i \leq 1 \). (Exposure is the proportion of the period of observation for which the policy has been in force.) We will be assuming that the \( t_i \) are known. If \( c_i \) is the number of claims in the period, and we assume a Poisson process with mean number of claims (per unit exposure time) \( \pi_i \) then \( c_i | t_i \sim \text{Po}(t_i \pi_i) \), \( P(c_i = 0 | t_i = 1) = e^{-\pi_i} \approx 1 - \pi_i \) and \( P(c_i = 0 | t_i) = e^{-t_i \pi_i} \approx 1 - t_i \pi_i \), provided \( t_i \pi_i \) is small. This gives

\[
f(w_i) = (\pi_i^*)^{w_i} (1 - \pi_i^*)^{1 - w_i}
\]

i.e. Bernoulli with \( \pi_i^* = t_i \pi_i \). We incorporate covariates through the logit link function on \( \pi_i^* \):

\[
\log \frac{\pi_i^*}{1 - \pi_i^*} = \eta_i
\]
i.e.
\[
\log \frac{\pi_i^* / t_i}{1 - \pi_i^* / t_i} = \eta_i
\] (2)
and the correction for differing periods of exposure enters the model through the modified link function (2). The predictor \( \eta_i \) is defined in the next section.

### 2.3 The mixture model
The zero-adjusted Inverse Gaussian (ZAIG) model is then
\[
f(y_i) = \begin{cases} 
1 - \pi_i^* & y_i = 0 \\
\pi_i^* \cdot \frac{1}{\sqrt{2\pi y_i \sigma_i}} \exp \left[ -\frac{1}{2} \left( \frac{y_i - \mu_i}{\mu_i \sigma_i} \right)^2 \right] & y_i > 0
\end{cases}
\]
which has \( E(y_i) = \pi_i^* \mu_i \) and \( \text{Var}(y_i) = \pi_i^* \mu_i^2 (1 - \pi_i^* + \mu_i \sigma_i^2) \). Following Rigby and Stasinopoulos (2005), who specify generalized additive models for the location, scale and shape parameters of a variety of distributions, we specify the following models on the parameters \( \mu_i, \sigma_i \) and \( \pi_i^* \):
\[
\begin{align*}
\log(\mu_i) &= x_{1\mu i}' \beta_\mu + f_\mu(x_{2\mu i}) \\
\log(\sigma_i) &= x_{1\sigma i}' \beta_\sigma + f_\sigma(x_{2\sigma i}) \\
\log \frac{\pi_i^* / t_i}{1 - \pi_i^* / t_i} &= x_{1\pi i}' \beta_\pi + f_\pi(x_{2\pi i})
\end{align*}
\]
where \( x_{1\mu i}, x_{2\mu i}, x_{1\sigma i}, x_{2\sigma i}, x_{1\pi i} \) and \( x_{2\pi i} \) are covariate vectors for \( \mu_i, \sigma_i \) and \( \pi_i^* \), which may be different, the same, or may have some but not all elements in common; \( \beta_\mu, \beta_\sigma \) and \( \beta_\pi \) are the corresponding parameter vectors; and \( f_\mu, f_\sigma \) and \( f_\pi \) are nonparametric functions, typically smoothing splines.

In order to correct for multiple claims in the period, we use the fact that, if \( y_j \sim IG(\mu, \sigma), \ j = 1, \ldots, c \) independently, then the total \( t = \sum_j y_j \) has the distribution
\[
t \sim IG(\mu^*, \sigma^*)
\]
where \( \mu^* = c\mu \) and \( \sigma^* = \sigma / c \). As
\[
\log(\mu^*) = \log(\mu) + \log(c)
\]
and
\[
\log(\sigma^*) = \log(\sigma) - \log(c)
\]
we use \( \log(c_i) \) and \( -\log(c_i) \) as offsets in the models for \( \mu_i \) and \( \sigma_i \) respectively, where \( c_i \) is the number of claims on policy \( i \). (A doubtful assumption here is that multiple claim amounts on the same policy are independent.)
3 Estimation

The ZAIG has been incorporated into the `gamlss` package in R (Stasinopoulos et al. (2006)). Maximum (penalised) likelihood estimation is used. The penalized log likelihood function of the model is maximized iteratively using either the RS or CG algorithm of Rigby and Stasinopoulos (2005), which in turn uses a back-fitting algorithm to perform each step of the Fisher scoring procedure. Both RS and CG algorithms use the log likelihood of the data, and its first derivatives (and optionally expected second derivatives) with respect to distributional parameters, which in this case are $\mu$, $\sigma$ and $\nu = \pi^*$. The CG algorithm, a generalization of the algorithm used by Cole and Green (1992), additionally uses the expected cross derivatives.

Motor vehicle insurance

The following covariates were available:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Characteristics of policy holder:</strong></td>
<td></td>
</tr>
<tr>
<td>Age band</td>
<td>1, 2, 3, 4, 5, 6 (1 is youngest)</td>
</tr>
<tr>
<td>Gender</td>
<td>male, female</td>
</tr>
<tr>
<td>Area of residence</td>
<td>A, B, C, D, E, F</td>
</tr>
<tr>
<td><strong>Characteristics of vehicle:</strong></td>
<td></td>
</tr>
<tr>
<td>Value</td>
<td>$0-$350,000</td>
</tr>
<tr>
<td>Make</td>
<td>A, B, C, D</td>
</tr>
<tr>
<td>Age</td>
<td>1, 2, 3, 4 (1 is recent)</td>
</tr>
<tr>
<td>Body type</td>
<td>bus, convertible, coupe, hatchback, hardtop, motorised caravan/combi, minibus, panel van, roadster, sedan, station wagon, truck, utility</td>
</tr>
</tbody>
</table>

Using the GAIC as model selection criterion, the following final model was selected:

\[
\log(\mu) = \text{age band} + \text{gender} + \text{area} + \text{offset}\{\log(\text{claims})\}
\]

\[
\log(\sigma) = \text{area} + \text{offset}\{-\log(\text{claims})\}
\]

\[
\log(\frac{\pi}{1-\pi}) = \text{age band} + \text{area} + \text{vehicle body type} + \text{spline(vehicle value)}
\]

Comments on the model

- **Model for $\pi$:** The model for the occurrence of a claim has terms for both policyholder and vehicle characteristics. Policyholder age, area and vehicle body are all categorical, so their form is not an issue; vehicle value is the only continuous covariate that we have, and it enters in the model in a smoothing spline form. This is understood when we examine the scatterplot of claim/no claim, with a smoothing spline, in Figure 2. The relationship is nonlinear; the probability of a claim is at a maximum for vehicle value around $40,000.
• Model for $\mu$: This contains only policyholder characteristics, which is surprising. A more complicated model involving vehicle value, make and some interaction terms, was a close second in the model selection. However, it was felt that this was too complex and difficult to interpret, so the simpler version was chosen.

• Model for $\sigma$: Area is the only covariate for $\sigma$. The variation of the claim size distribution with area is shown in Figure 3: it can be seen that areas D, E and F have shapes which are different from A, B and C, reflected in lower values for $\hat{\sigma}$. In fact areas D, E and F are rural whereas A, B and C are urban.

The explanatory variables age band and area appear in the model equations for both $\pi$ and $\mu$. It is of interest whether they affect the occurrence of a claim, and claim size, in the same way. Figure 4.a shows the effect of age band ($e^{\hat{\beta}}$), on both $\pi/(1-\pi)$ and $\mu$; figure 4.b shows the effect of area on both $\pi/(1-\pi)$ and $\mu$. Note that age band=3 and area=A are the reference categories. Age band 1 (the youngest drivers) increases both the odds of a claim and the mean claim size, to a similar extent; age bands 2 and 4 have a similar effect to age band 3; and age bands 5 and 6 (older drivers) decrease both the odds of a claim, and the mean claim size, their effect being greater on the odds of a claim. The effect of area on the odds of a claim, and mean claim size, is less clear: the only clear indication is that the mean claim size is increased in area F.
4 Conclusion

We introduce a method for modelling insurance claim sizes using a zero-adjusted Inverse Gaussian (ZAIG) model, which explicitly specifies a logit-linear model for the occurrence of a claim; and log-linear models for the mean claim size (given a claim has occurred); and the dispersion of claim sizes (given a claim has occurred). These three models may incorporate different covariates, or some of the same covariates, and may depend on common covariates in different ways. The Inverse Gaussian distribution accommodates the extreme right skewness of the claim distributions. Given the risk factors for a potential new policyholder, the expected claim size may easily be computed as the expected value of the ZAIG distribution, conditional on the covariate values; and quartiles of the claim size distribution may be calculated for each combination of covariate values. The ZAIG distribution introduced here is a useful distribution for modelling data where the total amount per unit of time is observed but where zero amounts are possible. Rainfall data and smoking/drinking habits data are possible candidates for modelling using the ZAIG distribution.

References

FIGURE 4. Effect of age category and area (exp($\beta$)) on occurrence of claim and claim size


A multivariate logistic regression model for randomized response data

Ardo van den Hout¹, Peter G.M. van der Heijden² and Robert Gilchrist³

¹ MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge, CB2 2SR, UK.
² Department of Methodology and Statistics, Faculty of Social Sciences, Utrecht University, P.O. box 80140, 3508 TC Utrecht, the Netherlands
³ STORM Research Center, London Metropolitan University, 2-16 Eden Grove, London N7 8EA, UK.

Abstract: A multivariate logistic regression model is discussed given that the dependent variables are subject to randomized response. Randomized response is an interview technique that can be used when sensitive questions have to be asked and respondents are reluctant to answer directly. The multivariate model is an adaptation of a model without randomized response variables and a Fisher scoring algorithm is used to maximize the likelihood. Randomized response data taken from a study into social benefit fraud are used to illustrate the approach.

Keywords: Multivariate logistic regression; misclassification; randomized response; sensitive questions.

1 Introduction

Randomized response (RR) is an interview technique that can be used when sensitive questions have to be asked and respondents are reluctant to answer directly ( Warner 1965; Chaudhuri and Mukerjee 1988). Examples of sensitive questions are questions about alcohol consumption, sexual behavior or fraud. RR variables can be seen as misclassified categorical variables where conditional misclassification probabilities are fixed by design. The misclassification protects the privacy of the individual respondent.

In recent years, RR techniques have been investigated and applied in the Netherlands. Van Gils, Van der Heijden, Laudy, and Ross (2003) report about rule transgression with respect to social benefits and Ellfers, Van der Heijden, and Hezemans (2003) used RR to study rule transgression regarding two Dutch instrumental laws. RR is also applied outside the Netherlands, see, e.g., Fisher, Kupferman, and Lesser (1992), and Lara, Strickler, Olavarrieta, and Ellertson (2004). A meta-analysis by Lensvelt-Mulder, Hox, Van der Heijden, and Maas (2005) shows that RR yields more valid prevalence estimates than other methods for sensitive questions.

234
In the case the sensitive question is a yes/no question, the univariate logistic regression model can be used to investigate how the sensitive behavior is associated with covariates such as, e.g., gender and age. Given the misclassification induced by the RR design, the standard logistic regression model has to be adapted to take into account the extra perturbation. For the Warner RR design, the univariate RR logistic regression model was first described by Maddala (1983). Further research in this area is presented by Van der Heijden and Van Gils (1996). In case all the covariates are discrete, loglinear models can be applied, see Chen (1989) and Van den Hout and Van der Heijden (2004).

Often in RR surveys there are several sensitive questions and we would like to be able to investigate - within one model - the associations between a number of RR variables and covariates. The present paper introduces a multivariate logistic regression model for dependent variables subject to RR. The model makes it possible to assess several RR variables and a set of covariates jointly. There are various ways to define a multivariate logistic regression model, see Fahrmeir and Tutz (2001, Section 3.5). The present paper extends the multivariate logistic regression model as presented by Glonek and McCullagh (1995) and shows how the model can be adapted to take the RR design into account. The estimation of this RR multivariate model is an extension of the Fisher scoring algorithm as presented by Glonek and McCullagh (1995).

The application at the end of the paper illustrates how the method can be used by analyzing RR data taken from a study into social benefit fraud.

2 The randomized response design

The forced response design (Boruch 1971) is an illustrative example of an RR design. Assume that the sensitive question asks for a yes or a no. After the sensitive question is asked, the respondent throws two dice and keeps the outcome hidden from the interviewer. If the outcome is 2, 3 or 4, the respondent answers yes. If the outcome is 5, 6, 7, 8, 9 or 10, the respondent answers according to the truth. If the outcome is 11 or 12, the respondent answers no. This design protects the privacy since an observed yes does not necessarily imply a latent yes.

Let \( Y \) be the latent binary variable that models the sensitive item, \( Y^* \) the binary variable that models the observed answer, and yes \( \equiv 1 \) and no \( \equiv 0 \). The RR design of the forced response design is given by

\[
P(Y^* = 1) = P(Y^* = 1|Y = 0)P(Y = 0) + P(Y^* = 1|Y = 1)P(Y = 1)
= \frac{2}{12}P(Y = 0) + \frac{11}{12}P(Y = 1)
\]

Note that probabilities \( P(Y^* = j|Y = k) \) are fixed for \( j, k \in \{0, 1\} \) by the known distribution of the sum of the two dice. RR variables can be seen as
misclassified variables, where conditional misclassification probabilities are
given by \( P(Y^* = j | Y = k) \).

The general RR design for variable \( Y \) can be described by using a matrix
\( P_Y = (p_{jk}) \), where \( p_{jk} = P(Y^* = j | Y = k) \). If the variable has 2 categories,
the misclassification design is given by \( \pi^* = P_Y \pi \), where \( \pi^* = (\pi_0^*, \pi_1^*)^t \) is
the distribution of the observed variable \( Y^* \) and \( \pi = (\pi_0, \pi_1)^t \) is
the distribution of the latent variable \( Y \). The advantage of the matrix nota-
tion becomes apparent when we assess the multivariate RR design. As an
example, if the misclassification of \( Y_1 \) is described by \( P_{Y_1} \) and the misclas-
sification of \( Y_2 \) is described by \( P_{Y_2} \), the misclassification of the Cartesian
product \( Y = (Y_1, Y_2) \) is described by \( P_Y = P_{Y_1} \otimes P_{Y_2} \), where \( \otimes \)
denotes the Kronecker product.

In RR surveys there is always the question whether respondents comply
with the design. An example of non-compliance would be a respondent
that answers no whereas the design forces the answer to be yes. We shall
not address this problem here although its importance is obvious. Instead
we refer to Cruyff, Van den Hout, Van der Heijden, and Böckenholt (2006)
- a paper in these proceedings - for further discussion and references.

3 Multivariate logistic regression

In the multivariate logistic regression model, the link between the linear
predictor \( \eta \) and the expected response vector \( \pi \) is
\[
\eta = C^t \log (L \pi),
\]
where \( C \) is the contrast matrix and \( L \) is the marginal indicator (Glonek
and McCullagh 1995).

We illustrate the model for the bivariate case with the binary responses \( Y_1 \)
and \( Y_2 \). Let \( \pi = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})^t \), where \( \pi_{kl} = P(Y_1 = k, Y_2 = l) \), for
\( k, l \in \{0, 1\} \). Given \( \eta = (0, \eta_{Y_1}, \eta_{Y_2}, \eta_{Y_1 Y_2})^t \), the link functions are given by
\[
\eta_{Y_1} = \log \frac{\pi_{1+}}{1 - \pi_{1+}}, \quad \eta_{Y_2} = \log \frac{\pi_{+1}}{1 - \pi_{+1}}, \quad \eta_{Y_1 Y_2} = \log \frac{\pi_{00} \pi_{11}}{\pi_{01} \pi_{10}},
\]
where the plus subscript denotes summation over the index. The regression
equations are given by
\[
\eta_{Y_1} = \beta_{Y_1}^t x_{Y_1}, \quad \eta_{Y_2} = \beta_{Y_2}^t x_{Y_2}, \quad \eta_{Y_1 Y_2} = \beta_{Y_1 Y_2}^t x_{Y_1 Y_2}.
\]

This model is a marginal model - it implies univariate logistic models for
both \( Y_1 \) and \( Y_2 \) marginally. The odds ratio is used to model the dependence
between \( Y_1 \) and \( Y_2 \). The contrast matrix and the marginal indicator for this
model are given by
\[
C^t = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 \\
\end{pmatrix}
\]
and
\[
L' = \begin{pmatrix}
1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1
\end{pmatrix}.
\]

The first element of \( \eta \) represents the null contrast \( \log (\pi^{++}) = 0 \). For an observation \( i \), the regression is given by \( \eta_i = X_i \beta \), where \( \beta = (\beta_{Y_1}, \beta_{Y_2}, \beta_{Y_1Y_2})^t \) and
\[
X_i = \begin{pmatrix}
0 & 0 & 0 & x_{Y_1,i} & 0 & 0 & 0 & x_{Y_2,i} & 0 \\
0 & x_{Y_1,i} & 0 & 0 & 0 & 0 & 0 & x_{Y_2,i} & 0
\end{pmatrix}.
\]

To estimate the RR multivariate logistic regression model, assume that data are given by independent observations \((y_i^*, X_i), i = 1, \ldots, n \), where \( y_i^* \) is a draw from a multinomial distribution with parameter \( \pi^* \). Let the RR design be described by \( P \). The kernel of the loglikelihood is
\[
l(\beta | \text{data}, P) = \sum_{i=1}^{n} (y_i^*)^t \log \pi^* = \sum_{i=1}^{n} (y_i^*)^t \log P \pi.
\]
Note that \( P \) will be the Kronecker product of the matrices that describe the RR design per RR question. For example, if there are two binary RR questions, \( P \) is a 4 \( \times \) 4 matrix, and \( y_i^* \in \{(1, 0, 0, 0)^t, (0, 1, 0, 0)^t, (0, 0, 1, 0)^t, (0, 0, 0, 1)^t\} \), for \( i = 1, \ldots, n \).

The maximization of the loglikelihood consists of two parts that are iterated. Given a starting value of \( \beta \), the first part uses a Newton-Raphson iteration to obtain \( \pi \) given \( \eta \). This part does not need an adaption for the RR design since the relation between the linear predictor and the latent distribution does not change. The second part is a Fisher scoring algorithm that takes the RR into account and updates the estimation of \( \beta \).

**Part one.** To invert equation (1), work with \( v = \log (\pi) \). The Newton-Raphson iterations are described by.
(a) Initial approximation is \( v_0 \).
(b) Next
\[
v_n = v_{n-1} - \left\{ \text{C}(\text{diag}(L \exp (v_{n-1})))^{-1} \text{C} \log (L \exp (v_{n-1})) - \eta \right\}^{-1} \text{C} \log (L \exp (v_{n-1})) - \eta.
\]
For details see Glonek and McCullagh (1995, Section 3).

**Part two.** The adaptation of the scoring algorithm is straightforward given the above. The derivative \( \partial \pi / \partial \beta \) is given by \( (C \text{diag}(L \pi)^{-1} L)^{-1} X \). The score statistic and the Fisher information matrix can be derived from the loglikelihood and are given by
\[
s(\beta | y_i^*) = \left( P \frac{\partial \pi}{\partial \beta} \right)^t \text{diag}(P \pi)^{-1} y_i^*
\]
\[
I(\beta | y_i^*) = \mathbb{E}_{Y_i^*} [s(\beta; y_i^*) s(\beta; y_i^*)]^t = \left( P \frac{\partial \pi}{\partial \beta} \right)^t \text{diag}(P \pi)^{-1} \left( P \frac{\partial \pi}{\partial \beta} \right).
\]
The information matrix is derived by using the fact that $Y^*_i$ is multinomially distributed with parameter vector $\pi^* = P\pi$. That is, $E_{Y^*_i} [Y^*_i] = \pi^*_i = E_{Y^*_i} [(Y^*_i)^2]$ and $E_{Y^*_i} [Y^*_i Y^*_k] = 0$ for $j \neq k$, see, e.g., Agresti (2002, Section 14.1).

Note that the above model does not make sense if the cross classification of the latent dependent variables contains one or more structural zeroes. For instance, in a $2 \times 2$ table with one structural zero, the estimated odds ratio is infinity. In the case of sample zeroes, we propose to smooth the data in the maximization procedure. That is, after part one add small probability mass to each entry of $\hat{\pi}$, normalize $\hat{\pi}$ such that the entries sum up to one, and next go to part two.

4 Application

We illustrate the above method by analyzing data taken from an RR survey concerning unemployment benefit in the Netherlands in 2004. Let $y^*_1$ denote the observed answer to the question whether the respondent applied frequently enough for new jobs (yes = 1, no = 0). Likewise let $y^*_2$ denote the observed answer to the question whether the respondent did any voluntary work without reporting this. In addition, $x_1$ is sex (male = 1, female = 0), $x_2$ is age in years, and $x_3$ denotes whether the income of the respondent constitutes the larger part of the income of the household (yes = 1, no = 0). We use the complete data and the sample size is $n = 753$.

The RR design is given by $p_{00} = 0.813$ and $p_{11} = 0.933$ and it is a slightly adapted form of the forced response design.

Table 1 presents the results of the analysis. Model 6 is defined by (1) and (2) and the regression equations $\eta_i = X_i \beta$ are given by $\beta = (\beta_{Y'_10}, \beta_{Y'_11}, \beta_{Y'_21}, \beta_{Y'_3}, \beta_{Y'_20}, \beta_{Y'_21}, \beta_{Y'_22}, \beta_{Y'_23}, \beta_{Y'_1Y'_20})^T$ and

$$X_i = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & x_{Y'_11} & x_{Y'_12} & x_{Y'_13} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & x_{Y'_21} & x_{Y'_22} & x_{Y'_23} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$ 

So the odd ratio is estimated by using one parameter, i.e., an intercept. Models 1 up to 5 are defined by restrictions on the parameters, see Table 1. The analysis of deviance would favor model 2 whereas the Akaike Information Criterion (AIC) would favor model 4. Going along with the AIC, we choose model 4 in which the regression coefficient of $x_3$ is restricted to be the same in the marginal models for $Y_1$ and $Y_2$, and $\beta_{Y'_11} = \beta_{Y'_21} = 0$.

Coefficient estimates for this model are given by

$$\hat{\beta}_{Y'_10} = -0.028 (0.544) \quad \hat{\beta}_{Y'_20} = -3.309 (0.683)$$
$$\hat{\beta}_{Y'_11} = -0.037 (0.014) \quad \hat{\beta}_{Y'_21} = 0.039 (0.013)$$
$$\hat{\beta}_{Y'_13} = \hat{\beta}_{Y'_23} = 0.444 (0.333) \quad \hat{\beta}_{Y'_1Y'_20} = 1.300 (0.536).$$
TABLE 1. Deviances of bivariate logistic regression models taken the RR design into account. Sample size is 753. Notation: $x_j(k,\ldots)$ means the covariate $x_j$ is included in the $k^{th}$ regression equation, where the order of the equations is given by (3).

<table>
<thead>
<tr>
<th>Model</th>
<th># parameters</th>
<th>Deviance</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intercept(1,2,3)</td>
<td>3</td>
<td>1928.033</td>
</tr>
<tr>
<td>2</td>
<td>Intercept(1,2,3) $x_2(1,2)$</td>
<td>5</td>
<td>1905.071</td>
</tr>
<tr>
<td>3</td>
<td>Intercept(1,2,3) $x_2(1,2), x_3(1)$</td>
<td>6</td>
<td>1903.799</td>
</tr>
<tr>
<td>4</td>
<td>Intercept(1,2,3) $x_2(1,2), x_3(1,2)$ and $\beta_{Y_13} = \beta_{Y_23}$</td>
<td>6</td>
<td>1901.638</td>
</tr>
<tr>
<td>5</td>
<td>Intercept(1,2,3) $x_2(1,2), x_3(1,2)$</td>
<td>7</td>
<td>1901.611</td>
</tr>
<tr>
<td>6</td>
<td>Intercept(1,2,3) $x_1(1,2), x_2(1,2), x_3(1,2)$</td>
<td>9</td>
<td>1900.889</td>
</tr>
</tbody>
</table>

where the estimated standard errors are within the brackets. The standard errors are estimated by evaluating the information matrix in the optimum.

The interpretation of the parameters of the marginal models for $Y_1$ and $Y_2$ is the same as in the univariate logistic regression model. One of the advantages of the multivariate model, however, is that we can easily investigate whether an effect of a covariate is the same in the marginal models. For the data at hand, this is illustrated by the chosen model 4 which states that the effect of whether or not the income of the respondent constitutes the larger part of the income of the household is the same for the prevalence of fraud regarding applying for jobs and the prevalence of fraud regarding voluntary work. More specific, model 4 states that a change from $x_3 = 0$ to $x_3 = 1$, means that the odds on fraud both on $Y_1$ and $Y_2$ changes multiplicatively by $\exp(\hat{\beta}_{Y_13}) = \exp(\hat{\beta}_{Y_23}) = \exp(0.444) = 1.559$. So the odds on fraud is higher when a person’s income constitutes the larger part of the income of the household. In addition, the multivariate model describes the association between the two dependent variables $Y_1$ and $Y_2$ by estimating the odds ratio. In model 4, the ratio of the odds in the two rows of the cross-classification of $Y_1$ and $Y_2$ is estimated to be $\exp(1.300) = 3.669$. This means that the odds of fraud on voluntary work given fraud regarding job applications is 3.669 times the odds of fraud on voluntary work given no fraud regarding job applications.
References


Prediction of risk of sepsis based on deviance in MGLM

Hrdličková Zuzana

1 Department of Applied Mathematics, Faculty of Science, Masaryk University, 602 00 Brno, Janáčkovo nám. 2a, Czech Republic, zuzka@math.muni.cz

Abstract: When analysing the risk of sepsis the multivariate generalized linear model (MGLM) can be employed. The paper describes an application of analysis of deviance to testing the influence of two studied polymorphism of genes. Power of the test based on deviance and its asymptotic approximation is also considered.

Keywords: MGLM; deviance; power; genetic prediction.

1 Introduction

The polymorphism of genes encoding protein receptors on immune system cells plays an important role in surviving sepsis. The aim of the grant of the GAČR N. 301/03/D196 ”Genetic prediction of the risk of sepsis in children” is to map the polymorphism of these genes and eventually find the combinations predisposing to severe sepsis in children hospitalized at the University Hospital Brno. This approach may lead to prediction of high-risk patients with sepsis so the effective treatment can be started early.

Previous analysis demonstrated that special attention should be paid to the genotypes of polymorphism IL6-176 and BPI-Taq.

2 Data and Model

Table 1 summarizes observed frequencies of children patients with different levels of sepsis and different variants of genes in the study. Due to zero observations, classes of genotypes were collapsed: variants 2, 3, 4 of IL6-176 correspond with genotypes aa, ab, bb, whereas variants 2, 3 of BPI-Taq correspond with genotypes aa and ”ab or bb”. Furthermore level 0 of sepsis indicates healthy patient (controls) and other levels correspond with an increasing severity of sepsis.

To analyse the data in Table 1 the two-way anova type MGLM [see Fahrmeir et al.] was applied. In two-way anova type MGLM we assume that q-dimensional vectors of response variables \( Y_{ij} = [Y_{ij1}, \ldots, Y_{ijq}]^T \) \( i = 1, \ldots, I \), \( j = 1, \ldots, J \) are independent and have regular densities from the exponential family. Furthermore, we assume that there exists a regular and injective link function \( g : \mathbb{R}^q \rightarrow \mathbb{R}^q \) of the expected values \( \mu_{ij} = E(Y_{ij}) \) for which

\[
g(\mu_{ij}) = \alpha + \beta_i + \gamma_j, \quad i = 1, \ldots, I, j = 1, \ldots, J,
\]
TABLE 1. Observed frequencies of patients with different levels of sepsis and variants of polymorphism IL6-176 and BPI-Taq.

<table>
<thead>
<tr>
<th>Polymorphism</th>
<th>Sepsis</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL6-176 (i)</td>
<td>BPI-Taq (j)</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>55</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>225</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>116</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>87</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>53</td>
</tr>
</tbody>
</table>

TABLE 2. Estimates of the parameters of model MGLM1.

<table>
<thead>
<tr>
<th>s</th>
<th>α</th>
<th>β_1</th>
<th>β_2</th>
<th>γ_j</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1.750</td>
<td>-0.390</td>
<td>-0.898</td>
<td>-0.731</td>
</tr>
<tr>
<td>2</td>
<td>-1.023</td>
<td>-0.204</td>
<td>-0.546</td>
<td>-0.344</td>
</tr>
<tr>
<td>3</td>
<td>-3.386</td>
<td>-0.575</td>
<td>0.118</td>
<td>-1.110</td>
</tr>
<tr>
<td>4</td>
<td>-1.584</td>
<td>-0.460</td>
<td>-1.023</td>
<td>-0.924</td>
</tr>
<tr>
<td>5</td>
<td>-2.588</td>
<td>-1.090</td>
<td>-0.855</td>
<td>-2.169</td>
</tr>
</tbody>
</table>

where α, β_1, ..., β_I, γ_1, ..., γ_J ∈ R^q are unknown vector parameters.

Note that there is only one response variable in each class of the model if the original data are grouped [see Fahrmeir et al.]. Thus the relative frequencies Y_{ijs} of having sepsis of level s = 0, ..., 5 can be described by multinomial distribution Mn_6(N_{ij}, \pi_{ij}), i = 2, 3, 4, j = 2, 3 in the studied case. Since the model with canonical link function was used and reference category of controls was chosen, we may write log odds in s-th category in the form

MGLM_1: \ln \frac{\pi_{ijs}}{\pi_{ij0}} = \alpha_s + \beta_s + \gamma_js, \quad s = 1, ..., 5

Maximum likelihood estimates of the parameters of MGLM_1 are displayed in Table 2. Remark, that additional constraints β_2 = γ_2 = 0 were imposed to obtain unique estimates of parameters. We may conclude that genotypes ab and bb of both IL6-176 and BPI-Taq decrease the relative risk π_{ijs}/π_{ij0} of sepsis of an arbitrary level s except for genotype bb of IL6-176 and sepsis level 3.

Analysis of deviance was carried out to examine the possibility of reduction of the two-way anova type MGLM to one-way anova type MGLMs

MGLM_2: \ln \frac{\pi_{ijs}}{\pi_{ij0}} = \alpha_s + \beta_{js}, \quad s = 1, ..., 5
TABLE 3. Analysis of Deviance.

<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>∆D</th>
<th>df</th>
<th>$\chi^2_{0.95}(\Delta df)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MGLM$_1$</td>
<td>9.280</td>
<td>10</td>
<td>18.307</td>
<td></td>
</tr>
<tr>
<td>MGLM$_2$</td>
<td>32.859</td>
<td>23.580</td>
<td>15</td>
<td>11.071</td>
</tr>
<tr>
<td>MGLM$_3$</td>
<td>26.702</td>
<td>17.422</td>
<td>20</td>
<td>18.307</td>
</tr>
</tbody>
</table>

and

$$MGLM_3 : \ln \frac{\pi_{ij}}{\pi_{ij0}} = \alpha_s + \gamma_{js}, \quad s = 1, \ldots, 5$$

The results of the analysis of deviance, summarized in the Table 3, suggest rejecting the hypothesis that MGLM$_1$ can be reduced to MGLM$_2$ in contrast to the hypothesis of reduction of MGLM$_1$ to model MGLM$_3$.

3 Power of Test based on Deviance

The test of the hypothesis of reduction of MGLM$_1$ to MGLM$_2$, can be rewritten as test of the hypothesis $H_0 : \gamma_3 = 0$ against the alternative $A_1 : \gamma_3 = \gamma_0$, where $\gamma_0 = [-0.731, -0.344, -1.110, -0.924, -2.169]^T$. If alternatives $A(c) : \gamma_3 = c\gamma_0$, $c > 0$ are specified, the power of the test based on deviance at the significance level $\alpha$ can be written as $\beta_{\alpha,c} = P(\Delta D > \chi^2_{1-\alpha}(\Delta df)|A(c))$. Since the power can not be expressed analytically, it’s asymptotic approximation obtained by generalizing the approach of [Cordeiro et al. 1994] can be used. Using this technique, an approximation of the distribution function of statistics $\Delta D$ is obtained in the form

$$F(t) = G_{p-r,\lambda}(t) + \sum_{k=0}^{2} b_k G_{p-r+2k,\lambda}(t) + o(n^{-1/2}),$$

where $G_{p-r,\lambda}$ is the distribution function of a noncentral $\chi^2$ distribution with $p-r$ degrees of freedom and parameter of noncentrality $\lambda$, coefficients $b_0, b_1, b_2$ and $\lambda$ are functions of $c\gamma_0$ and cumulants of the derivatives of the loglikelihood function both appropriate to the simultaneous density of vectors $Y_{ij}, i = 2, 3, 4, j = 2, 3, \{Hrdličková\}$.

Figure 1 displays the comparison of the simulated and approximated power of the test based on deviance corresponding to the data in Table 1 and estimates of the parameters of MGLM$_1$ in Table 2 at type I error 0.05. Then, for example, it can be seen that if type I and type II error was set to 0.05 and 0.1 respectively, the alternative would be distinguished even for $c = 0.55$. 


4 Conclusions

Dependence of a level of sepsis on variants of polymorphism IL6-176 and BPI-Taq was studied. It appears that genotypes ab and bb of both IL6-176 and BPI-Taq mostly decrease the relative risk of sepsis of an arbitrary level $s$. The power of the test was examined to show, that the alternatives could be distinguished even for $c < 1$ while analysing the influence of BPI-Taq at the observed number of patients in each class of two-way anova type MGLM.

Acknowledgments: The research was supported by research proposal MSM0021622418, by grant of GAČR N. 201/05/H007 and in part by grant of GAČR N. 301/03/D196. The author is grateful to Jaroslav Michálek for providing the data.

References


Modelling of recurrence of adenomas: a latent variable approach

Chiu-Hsieh Hsu

1 Mel and Enid Zuckerman College of Public Health and Arizona Cancer Center, University of Arizona, Tucson, AZ 85724, USA, phsu@azcc.arizona.edu

Abstract: We treat the number of recurrent adenomas as a latent variable and then use a mixture distribution to model the number of observed recurrent adenomas. This approach is equivalent to zero-inflated Poisson regression, which is a method used to analyze count data with excess zeros. In many cancer studies, patients often have variable follow-up. When the disease of interest is subject to late onset, ignoring the length of follow-up will underestimate recurrence rate. We modify zero-inflated Poisson regression through a weight function to incorporate the length of follow-up into analysis. We motivate, develop, and illustrate the methods described here with an example from a colon cancer study.

Keywords: Colon cancer; latent variable; weight function; zero-inflated Poisson.

1 Introduction

Colorectal cancer is one of the leading causes of cancer death. Many studies using recurrence of colorectal adenomas as a surrogate endpoint have been launched to develop prevention strategies for this serious disease. However, the adenoma data are often subject to measurement error due to misclassification at colonoscopy, especially for diminutive polyps. Most of adenoma datasets contain a high percentage of subjects with no adenomas. This could result from the misclassification of small adenomas at colonoscopy. In addition, some participants did not adhere strictly to the schedule of examinations for colonoscopy. Due to this, a patient with no adenomas at one year should be treated differently from a patient with no adenomas at three years for estimation of recurrence rate at the end of study. When the disease of interest is subject to late onset, the methods that do not make use of the length of follow-up will underestimate recurrence rate.

In this paper, we employ a latent variable to develop an approach, which accounts for measurement error and excess zeros, to estimating recurrence rate. This approach is equivalent to zero-inflated Poisson (ZIP) regression (Lambert, 1992). To adjust for variable follow-up, we adapt the idea in Cheung and Chappell (2000) to modify ZIP regression by incorporating a weight function, which is a function of the length of follow-up, into analysis of the adenoma.
2 Methods

2.1 A Latent Variable Model

Suppose that there are \( n \) subjects in a study. For each subject \( i = 1, \ldots, n \), there are \( Y_i \) observable recurrent adenomas. Let \( X_i \), an unobservable latent variable, denote the number of true recurrent adenomas for subject \( i \). In most situations, an adenoma might go undetected, but not over-counted. We, therefore, assume that \( \Pr(Y_i = 0 | X_i = 0) = 1 \) and \([Y_i | X_i > 0]\) is from a Poisson distribution with parameter \( \lambda \), and \( \Delta_i = I(Y_i > 0) \), where \( I(\cdot) \) is an indicator function. We assume that these \( n \) subjects come from a random sample. Let \( p = \Pr(X_i > 0) \) denote recurrence rate. The marginal distribution of \( Y_i \) can be expressed as

\[
\Pr(Y_i = y_i; p, \lambda) = \begin{cases} 
(1 - p) + pe^{-\lambda}, & \text{if } y_i = 0, \\
pe^{-\lambda} \frac{\lambda^{y_i}}{y_i!}, & \text{if } y_i > 0.
\end{cases}
\]

This corresponds to a ZIP model where \( p \) presents the recurrence rate of adenomas and under independent measurement error assumptions, \( \lambda \) provides the information of disease progression given recurrence of adenomas.

2.2 Time-Adjusted Recurrence Rate

Let \( t \) denote the follow-up time. Based on the ideas in Cheung and Chappell (2000), we use an increasing weight function, \( w(t) \), where \( 0 \leq w(t) \leq 1 \), to incorporate the length of follow-up into estimation of recurrence rate. If \( w(t) = 1 \), it indicates that the patient’s information is fully observed. Otherwise, the information is partially observed. This assumes that a patient with a longer follow-up provides more information about recurrence rate compared to a patient with a shorter follow-up. Based on this approach, for patient \( i \) the probability of no recurrent adenomas at \( t_i \), i.e. \( \Pr(X_i = 0; t_i) \), is defined as \( 1 - F(t_i) \), where \( F(\cdot) \) is the distribution function of time to recurrence. In this paper we assume that \( F(t_i) \), i.e. recurrence rate at \( t_i \) \((\Pr(X_i > 0; t_i))\), is equal to \( w(t_i) \Pr(X_i > 0) \). The distribution of \( Y_i \) given \( t_i \) follow-up can be expressed as

\[
\Pr(Y_i = y_i; t_i, p, \lambda) = \begin{cases} 
1 - w(t_i)p + w(t_i)pe^{-\lambda}, & \text{if } y_i = 0, \\
w(t_i)p \frac{\lambda^{y_i}}{y_i!}, & \text{if } y_i > 0.
\end{cases}
\]

The ZIP incorporates the length of follow-up by using a weight function of “time-to-event” denoted as W-ZIP. Assume that the parameter in the weight function is distinctive of the parameters in the ZIP model. The weight function will only be involved in the term with censored data. This implies that the \( w(t) \) is equivalent to 1 for the patients who have adenomas when estimating the parameters and, therefore, will not impact estimation of the disease progression. The weight function can be obtained by fitting a Kaplan-Meier curve to the observed time to recurrence data (denoted by W-ZIP\(^{KM}\)).
3 Application to Wheat Bran Fiber (WBF) Study

In the WBF study, 1429 men and women were randomly assigned to one of the two dietary groups, low fiber and high fiber. Of 1429 subjects, there were 1303 subjects who completed the study, 584 in the low-fiber group and 719 in the high-fiber group. Initially, the last colonoscopy was planned to be conducted at three years after randomization. However, some patients went through their final colonoscopy before the planned time. The median times from randomization to the last follow-up colonoscopy were 36 months in the low-fiber group and 34 months in the high-fiber group. By the time of the final colonoscopy, 299 (51.2%) subjects in the low-fiber group were identified with at least one adenoma and 338 (47.0%) subjects in the high-fiber group were identified with at least one adenoma. The maximum follow-up, $\tau$, is set at three years to calculate the weight for the W-ZIP method. The results are provided in Table 1. For recurrence rate, after adjusting for possible confounders the ZIP method indicates that the high fiber group has a significantly lower recurrence rate compared to the low fiber group with an OR 0.67 (95% CI: 0.48 to 0.93; p-value=0.02). The W-ZIP$^{KM}$ method also yields a significant treatment effect with an OR 0.78 (95% CI: 0.64-0.95; p-value=0.02). Whereas, the PO method, which simply applies logistic regression to the observed recurrent data, indicates there is no significant treatment effect. The results for disease progression can also be found in Table 1. In a summary, the ZIP and W-ZIP methods could provide an adjustment for measurement error and excess zeros. In addition, the W-ZIP method also has a potential to adjust for the late onset of adenomas.

4 Simulation Study

We perform a simulation study to investigate the properties of the WZIP method under a variety of “time-to-event” distributions. We mainly focus on the situation without any covariates, which is aimed at comparing the estimate of the recurrence rate between the Kaplan-Meier and WZIP methods in this paper. Each of 1000 datasets was generated as follows. The time to recurrence of adenomatous polyps ($T$) is from a distribution function $F(t)$ and the time to the final colonoscopy ($C$), the follow-up time, is from a distribution function $G(c)$. To mimic the WBF study, $G(c)$ is chosen to have the median follow-up at 3 years and the maximum follow-up is set at 3/5 years, i.e. $C^* = \min(3/5, C)$. In the situation that $\min(T, C^*) = C^*$, $Y$, the observed recurrent adenomatous polyps, is equal to zero with probability one. In the situation that $\min(T, C^*) = T$, $Y$ is generated from a Poisson distribution with parameter $\lambda$. For each of the datasets, there are 100 subjects.

The results are provided in Table 2. As expected, the biases of the ZIP method are consistently greater than that of WZIP method in all situations.
The bias results in low coverage rates for the ZIP method. The performance of the Kaplan-Meier method highly depends on $Pr(Y = 0) = e^{-\lambda}$. As $\lambda$ decreases, the bias of the KM method significantly increases. Overall, the results in Table 3 show that the WZIP method in which the weight is derived from a nonparametric survival function can provide reasonable estimates for recurrence rate and is robust to the assumption of the underlying distribution of time-to-recurrence. The WZIP method can be considered as a robust weighted-ZIP method.

5 Discussion

The research in this paper treats the number of recurrent adenomas as a latent variable and then uses a weight function to modify ZIP regression to incorporate the length of follow-up into analysis. We realize that the performance of the modified ZIP method highly depends on the selection of the weight function. We show in Hsu (2005) that a nonparametric weight function derived from a biased Kaplan-Meier survival curve (the bias is due to the measurement error on censored indicator) can produce comparable point estimates to the true value and doesn’t rely on any parametric assumptions. The weighted ZIP method using a nonparametric weight function, i.e. $W$-ZIP$^{KM}$, to incorporate the length of follow-up into analysis is robust to the assumption of the underlying distribution for time to recurrence. The research in this paper implicitly assumes that time to recurrence is independent of follow-up time, i.e. “independent censoring”. However, if the reason for changing the examination schedule for patient’s final colonoscopy is related to their disease status, the independent censoring assumption will be violated. Future research can focus on developing strategies that can deal with both measurement error and dependent censoring and also can jointly model the disease progression.

References


TABLE 1. Analysis of the WBF data

<table>
<thead>
<tr>
<th>Variable</th>
<th>Logistic Regression</th>
<th>Poisson Regression**</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est</td>
<td>SE</td>
</tr>
<tr>
<td>Treatment</td>
<td>-0.138</td>
<td>0.123 (0.26)</td>
</tr>
<tr>
<td># of baseline adenomas</td>
<td>0.345</td>
<td>0.053 (0.00)</td>
</tr>
<tr>
<td>Sex</td>
<td>0.429</td>
<td>0.127 (0.00)</td>
</tr>
<tr>
<td># of colonoscopy</td>
<td>0.757</td>
<td>0.081 (0.00)</td>
</tr>
<tr>
<td>Randomization scheme</td>
<td>-0.091</td>
<td>0.153 (0.55)</td>
</tr>
</tbody>
</table>

Recurrence

<table>
<thead>
<tr>
<th>ZIP</th>
<th>W-ZIP*3.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>-0.404</td>
</tr>
<tr>
<td># of baseline adenomas</td>
<td>0.308</td>
</tr>
<tr>
<td>Sex</td>
<td>0.464</td>
</tr>
<tr>
<td># of colonoscopy</td>
<td>0.588</td>
</tr>
<tr>
<td>Randomization Scheme</td>
<td>-0.162</td>
</tr>
</tbody>
</table>

Progression

<table>
<thead>
<tr>
<th>ZIP</th>
<th>W-ZIP*3.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>0.274</td>
</tr>
<tr>
<td># of baseline adenomas</td>
<td>0.146</td>
</tr>
<tr>
<td>Sex</td>
<td>0.077</td>
</tr>
<tr>
<td># of colonoscopy</td>
<td>0.419</td>
</tr>
<tr>
<td>Randomization scheme</td>
<td>0.089</td>
</tr>
</tbody>
</table>

*p-value inside the parenthesis.

** treat follow-up time (years) as an offset.
### TABLE 2. Monte Carlo Results: Estimation of recurrence of adenomatous polyps at 3/5 years. Sample size 100 and \( G(t) \sim \text{Exp}(0.23) \).

<table>
<thead>
<tr>
<th>Method</th>
<th>( p ) bias</th>
<th>SD(^1)</th>
<th>SE(^2)</th>
<th>CR(^3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUE ( p = \text{F}(3) = 0.362 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KM</td>
<td>0.349 -0.013</td>
<td>0.0576</td>
<td>0.0591</td>
<td>81.9</td>
</tr>
<tr>
<td>ZIP</td>
<td>0.270 -0.092</td>
<td>0.0453</td>
<td>0.0461</td>
<td>48.8</td>
</tr>
<tr>
<td>WZIP(^K,M)</td>
<td>0.369 0.007</td>
<td>0.0609</td>
<td>0.0605</td>
<td>94.7</td>
</tr>
<tr>
<td>F(t)~\text{Exp}(0.15); ( Y \sim \text{Poisson}(3.0) )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KM</td>
<td>0.245 -0.117</td>
<td>0.0554</td>
<td>0.0470</td>
<td>35.8</td>
</tr>
<tr>
<td>ZIP</td>
<td>0.275 -0.087</td>
<td>0.0822</td>
<td>0.0827</td>
<td>67.9</td>
</tr>
<tr>
<td>WZIP(^K,M)</td>
<td>0.404 0.042</td>
<td>0.1222</td>
<td>0.1197</td>
<td>95.4</td>
</tr>
<tr>
<td>F(t)~\text{Exp}(0.15); ( Y \sim \text{Poisson}(1.0) )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KM</td>
<td>0.605 -0.026</td>
<td>0.0627</td>
<td>0.0823</td>
<td>86.6</td>
</tr>
<tr>
<td>ZIP</td>
<td>0.463 -0.168</td>
<td>0.0435</td>
<td>0.0525</td>
<td>6.0</td>
</tr>
<tr>
<td>WZIP(^K,M)</td>
<td>0.636 0.005</td>
<td>0.0650</td>
<td>0.0658</td>
<td>94.9</td>
</tr>
<tr>
<td>TRUE ( p = \text{F}(5) = 0.631 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F(t)~\text{Weibull}(0.2,0.6); ( Y \sim \text{Poisson}(3.0) )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KM</td>
<td>0.526 -0.105</td>
<td>0.0748</td>
<td>0.0755</td>
<td>60.9</td>
</tr>
<tr>
<td>ZIP</td>
<td>0.452 -0.179</td>
<td>0.0709</td>
<td>0.0691</td>
<td>31.9</td>
</tr>
<tr>
<td>WZIP(^K,M)</td>
<td>0.677 0.046</td>
<td>0.1046</td>
<td>0.0970</td>
<td>91.8</td>
</tr>
</tbody>
</table>

\(^1\) empirical standard deviation of 1000 point estimates.
\(^2\) average of 1000 estimated standard errors.
\(^3\) fraction of 95% confidence intervals which contain the true value based on 1000 datasets.
Simultaneous analysis of variables observed in a series of trials

Zygmunt Kaczmarek¹, Elżbieta Adamska¹ and Teresa Cegielska-Taras²

¹ Institute of Plant Genetics, Polish Academy of Sciences 60-479 Poznań, ul. Strzeszyńska 34, Poland, zkac@igr.poznan.pl
² Plant Breeding and Acclimatization Institute, Poznań, Poland.

1 Introduction

Successful new genotypes, in particular doubled haploid lines (DH lines), must show high performance for various characters. However, their superiority should be reliable over a wide range of environmental conditions. The basic cause of differences between genotypes in their stability is the wide occurrence of genotype-environment interactions (GE interactions). Methods of statistical analysis of a series of plant breeding experiments, developed by Caliński et al. (1997, 2005), allow to evaluate individual genotypes in various environmental conditions including assessment of stability and adaptability.

The aim of this paper is to show how to use MANOVA and other statistical methods for estimation rapeseed DH lines and testing the hypotheses interesting from the point of view their stability, with respect to some function of favourable fatty acids important for industrial purposes. The quality of plant is determined by its component fatty acids. Thus oilseed *Brassica napus* DH lines with high oleic and low linolenic acid contents are highly desirable for biofuel and lubricant production.

2 Description of the data

The data analysed in the paper come from experiments, in which thirty seven genotypes of winter oilseed rape (32 doubled haploids lines, 2 parental forms P₁ and P₂, 2 segregating populations F₂ and F₃ and standard variety *Kana*) were studied in six environments, denoted here as A, B, C, D, E, F. The experiments were carried out in a randomized complete block design with three replications. The composition of fatty acids in seed with the use of gas chromatography method was estimated. The contents of five fatty acids: palmitic (C16:0), stearic (C18:0), oleic (C18:1), linoleic (C18:2)
and linolenic (C18:3) were denoted. Because the main aim of this series of experiments was the evaluation of DH lines as regards their further use in some branches of industry we looked for these lines of rapeseed whose oil is characterized by desirable composition of saturated and unsaturated fatty acids. The calculations were made for all 37 genotypes. It would not be possible to present them here in full. Only some results of the analysis of 12 randomly chosen DH lines will be shown. In fact these DH lines form only a subset of a much larger set of genotypes compared in all the trials.

3 Model of observations

The methods for the analysis of GE interactions and their structure with reference to the environments and to the genotypes, given by Kaczmarek (1986) and Caliński et al. (1997, 2005) were used. Considering each environment as a variable and genotypes as objects, the multivariate techniques might be used to describe the performance of genotypes. In order to present the methods of stability analysis the multivariate model for a random vector $y_j$ of the observed trials means for all $I$ genotypes at environment $j$ ($j = 1, 2, ..., J$) can be written as follows:

$$y_j = \mu + \alpha(j) + a^E(j) + e_j,$$

where $y_j = [y_{1j}, y_{2j}, ..., y_{Ij}]'$, $\mu = [\mu_1, \mu_2, ..., \mu_I]'$ in which $\mu_i = \mu_i(\cdot)$ denotes the average capacity of genotype $i$ ($i = 1, 2, ..., I$) over all environments, $\alpha(j) = [\alpha_1(j), \alpha_2(j), ..., \alpha_I(j)]'$ in which $\alpha_i(j) = \mu_i(j) - \mu_i(\cdot)$ denotes the average deviation of the capacity of genotype $i$ in environment $j$, $a^E(j) = [a^E_1(j), a^E_2(j), ..., a^E_I(j)]'$ in which $a^E_i(j)$ denotes a deviation under the environment of the site of the experiment at location $j$, and $e_j = [e_{1j}, e_{2j}, ..., e_{Ij}]'$, in which $e_{ij}$ is the average error for genotype $i$ at environment $j$ from the experiment. Since at each location usually not more than one experiment is carried out, the distinction between $\alpha(j)$ and $a^E(j)$ would not be possible in the analysis. Therefore, assuming that $a^E_i(j) = 0$ for all $i$ and $j$, model (1) is as follows:

$$y_j = \mu + a^E(j) + e_j.$$  

(2)

To apply MANOVA methods to model (2) it is assumed that the all random vectors $y_j$ are all independent, each of an $I$-variate normal distribution

$$y_j \sim N_I(\mu, \Sigma_y)$$

(3)

with the dispersion matrix $\Sigma_y$.

However, to allow for comparisons among the genotypes, the model (2) needs a further transformation, to $z_j = Gy_j$, for any $j$, where

$$G = I_I - I^{-1}_I 1_I 1_I',$$
with $1_I$ denoting the $I \times I$ vector of unit elements. The transformed model is then

$$z_j = \alpha^G + \alpha^{GE}(j) + f_j,$$

(4)

where the vector $\alpha^G = G\mu$ is composed of the genotype main effects, $\alpha^{GE}(j) = Ga^{E}(j)$ of the genotype interactions with the environment of the site of the experiment at location $j$, and finally $f_j = Ge_j$ is composed of the genotype error deviations from the average experimental error.

From the assumptions for (3), the random vectors $z_j$ are distributed independently, of the form

$$z_j \sim N_I(\alpha^G, \Sigma_z),$$

(5)

with the dispersion matrix $\Sigma_z$.

The assumption that the vectors (2) and (4) are distributed independently is reasonable, while the assumption of normality is a convenient simplification.

The adopted model with the distributions (3) and (5), implies the use of MANOVA as the main technique of analysing the data. The relevant sum of squares and products matrices and the least squares estimates of the parameters were obtainable by Kaczmarek (1986) and Caliński et al. (1997).

To test significance of the various sources of variation considered in the model, the appropriate ANOVA is performed. The multivariate character of the data implies the use of MANOVA for this purpose. For testing the hypotheses of no main effects of the environments (E) and no genotype-environment (GE) interactions, the usual mean square ratios are used as test statistics, with the error mean square (obtained by pooling the error sums of squares from individual experiments) as denominator. Information concerning each genotype separately can be received by the use $F$-statistics for main effects and for interactions.

## 4 Results

Statistical calculation of the data described in Section 2 were made by the computer program SERGEN (Caliński et al., 1994). As mentioned in this section the data come from the experiment in which 37 genotypes of winter rapeseed were denoted with respect to five fatty acids. For the estimation of DH lines in reference to their further utilisation it is necessary to impose certain requirements considering the content and proposition of fatty acids (Adamska et al., 2004). In particular the preferable fatty acids composition for production of biofuel and lubricant has the highest level of oleic acid (C18:1), as well as lowest amount of linolenic acid (C18:3). In order to estimate DH lines with respect the purpose mentioned above, it may be reasonable to transform initial five variables into new one, C18:1 – C18:3. It should be noticed that using interpretation for new variable, DH lines
should be the best when the values of this variable are the highest. Table 1 shows, for each genotype the main effect estimate, the $F$-statistic values for it and the $F$-statistic value for interaction and regression of interaction effects on environment.

**TABLE 1.** Estimates and results of testing the hypotheses concerning DH lines in respect of simultaneous oleic and linolenic acid contents

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Means for acids</th>
<th>Difference of acids : oleic – linolenic (C18:1-C18:3)</th>
<th>$F$-statistic for main effect</th>
<th>Regression on environment GE interaction</th>
<th>$F$-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>oleic</td>
<td>linolenic</td>
<td>estimate of main effect</td>
<td></td>
<td>coefficient</td>
</tr>
<tr>
<td>DH1</td>
<td>62.3</td>
<td>10.2</td>
<td>0.74</td>
<td>49.91</td>
<td>0.17</td>
</tr>
<tr>
<td>DH2</td>
<td>64.0</td>
<td>9.8</td>
<td>2.90</td>
<td>20.16</td>
<td>6.39</td>
</tr>
<tr>
<td>DH3</td>
<td>64.1</td>
<td>10.0</td>
<td>2.77</td>
<td>27.32</td>
<td>4.30</td>
</tr>
<tr>
<td>DH7</td>
<td>62.3</td>
<td>10.6</td>
<td>0.40</td>
<td>1.89</td>
<td>1.30</td>
</tr>
<tr>
<td>DH11</td>
<td>66.2</td>
<td>8.4</td>
<td>6.49</td>
<td>116.07</td>
<td>5.56</td>
</tr>
<tr>
<td>DH14</td>
<td>59.0</td>
<td>12.0</td>
<td>-4.24</td>
<td>59.12</td>
<td>4.66</td>
</tr>
<tr>
<td>DH19</td>
<td>59.8</td>
<td>11.3</td>
<td>-2.81</td>
<td>29.17</td>
<td>4.14</td>
</tr>
<tr>
<td>DH22</td>
<td>62.7</td>
<td>9.4</td>
<td>1.99</td>
<td>5.94</td>
<td>10.16</td>
</tr>
<tr>
<td>DH26</td>
<td>62.8</td>
<td>9.7</td>
<td>1.87</td>
<td>14.73</td>
<td>3.64</td>
</tr>
<tr>
<td>DH28</td>
<td>60.2</td>
<td>11.1</td>
<td>-2.14</td>
<td>41.96</td>
<td>1.67</td>
</tr>
<tr>
<td>$P_1$</td>
<td>63.3</td>
<td>9.9</td>
<td>2.07</td>
<td>58.68</td>
<td>1.12</td>
</tr>
<tr>
<td>Kana</td>
<td>62.6</td>
<td>9.7</td>
<td>1.56</td>
<td>47.91</td>
<td>0.78</td>
</tr>
<tr>
<td>Mean</td>
<td>61.7</td>
<td>10.4</td>
<td>Critical values: $F_{0.05}$</td>
<td>6.61</td>
<td>2.23</td>
</tr>
</tbody>
</table>

The analysis of the data allowed to find genotypes stable ($P_1$, Kana, DH1) and unstable (DH11, DH2, DH3, DH26) with high oleic and simultaneously low linolenic acid contents.

5 Conclusion

Breeding programmes should be planned in such a way that selection of new DH lines with the stability and desired adaptability to environmental conditions should be possible to the same degree as selection based on the phenotypic performance of desirable traits.


Analysis of variables observed in a series of trials


A latent variable model for Australian student’s creativity

Helen Kandilorou\(^1\), Zoe Georganta\(^2\) and Alexandra Livada\(^1\)

\(^1\) Department of Statistics, Athens University of Economics and Business, 76 Patission str, 10434 Athens, Greece
\(^2\) Department of Applied Informatics, University of Macedonia, 156 Egnatia Ave, Thessaloniki 54006, Greece

Abstract: The statistical analysis of results of a series of trials conducted with the same set of genotypes (DH lines) and repeated at several environments is described. The analysis is based on a Scheffé-type mixed model for observations. A special attention is given to estimation and hypothesis testing problems concerning the genotype-environment interactions (GE interactions). Application of this approach was shown on an example concerning the evaluation of DH lines of oilseed rape with respect to some function of favourable fatty acids important for industrial purposes.

Keywords: Creativity; latent variables; tertiary education.

1 Introduction

Creativity can be defined in many different ways. There is no one definition of creativity that everyone can agree with. So, it turns up that creativity is difficult to be defined, although it is particularly easy to be recognized. According to Amabile (1988, 1996), creativity is the production of novel and useful ideas by an individual or small group of individuals working together. That is, creativity requires a collaborative environment.

Creativity is not “a flash of inspiration out of the blue”, but it relates a concept to a particular body of knowledge, which is as “vital as the novel idea and really creative people spend years and years acquiring and refining their knowledge base—be it music, mathematics, arts, sculpture or design” (http://iexchange.London.edu). Creativity cannot be ordered, but it relies heavily on intrinsic motivation (Amabile, 1996, and Amabile, et al. 1996) and can be stimulated and supported through training and education. This paper does not aim at developing any new theory of creativity. It’s given that there are societal characteristics, even in western societies, which tend to stifle creative initiatives especially in young generations. The ultimate purpose of this survey is to explore some societal characteristics, which influence the student’s creativity, in the Faculty of Economics and Business at the University of Sydney. These factors are: social compromise,
parents’ social-status and psychological oppression within family (the psychological pressure exercised by the family concerning their personal believes about the children’s educational orientation), as well as student’s stress development due to the lack of learning skills.

2 The econometric creativity model

In this paper we construct a latent variable model to analyze the direct and indirect relationships between social compromise (COMPR), student’s creativity (CREAT), parents’ social-status (PARSTAT), psychological oppression within family (PARREP) and student’s stress development due to the lack of learning skills (STRESS). The following figure shows the relationships involved. In Figure 1, two sets of variables are distinguished, namely the observed indicator variables compiled on the basis of information collected in the authors’ statistical survey, and the non-observed latent or conceptual variables, which are denoted by stars. The indicator variables, which are observed with systematic measurement errors, are used as proxies for the conceptual variables, which are the “true” or “without errors” variables.

3 The data set and the variables

The empirical investigation is based on a statistical survey, which was conducted by the authors. A sample of 263 students was randomly selected from the population. The survey lasted three weeks in October-November 2004. The questionnaire along with a letter was emailed to the selected students, on which confidentiality issues and the purpose of the survey were discussed and explained. The response rate of this survey was 76%, which means 200 questionnaires were completed and emailed back to the authors. The completed questionnaires were used to construct indicator variables for parents’ social-status, psychological oppression within family, student’s stress, social compromise, and student’s creativity. PARSTAT denotes parents’ social status, measured as the average of both parents’ education.
H. Kandilorou et al. 259

PARREP denotes psychological oppression within family, indicating whether parents actively encourage their child to pursue a more preferable career, opting out of university. STRESS indicates the average student’s stress development due to the lack of learning skills (writing, committing information from lectures to long-term memory, and book reading strategy), which is based on three questions of the questionnaire; in each question, student’s stress has eight different values (value 1 reflects “no student’s stress” and value 8 reflects a severe student’s stress). CREAT1 and CREAT2 are two indicators of student’s creativity. Three different values are assigned to CREAT1, according to the way a student takes lecture-notes; value 3 and 1 reflect the most and least creative way, correspondingly. CREAT2 reaps also three different values, reflecting mutually the respondent’s stated opinion on two questions (value 3 and 1 are a sign of the most and least creative way of thinking respectively). CREAT* denotes the conceptual variable of creativity. COMPR* is also a conceptual variable represented by the two indicator-variables, COMPR1 and COMPR2, which are two indices constructed to measure social compromise in two different motivation areas. COMPR1 takes two different values, 0 and 1, where value 1 reveals social compromise when a respondent denotes that he/she would like to opt out university, if any chance, in favor of another preferable career, whereas value 0 implies no social compromise. COMPR2 takes three different values; value 3 manifests the highest level of social compromise (whether the student would resist and be uncooperative if his/her child decided to give up university studies for a new desired career) and value 1 shows the lowest level. The above indicator variables were used to estimate the empirical model.

4 The results

According to Table 1, the fit of both models (CASE I and II) is satisfactory as can be seen from both values of Minimum-Fit-Function chi-square (MFF $\chi^2$) and Normal-Theory-Weighted-Least-Squares chi-square (NTWLS $\chi^2$). Moreover, the estimated variance-covariance matrix of the exogenous indicator variables is very close to the observed variances and covariances, giving an additional indication of the good fit of the model as a whole. The effect of social compromise on creativity (Table 1) is negative and significant, regardless of what the utilized sample is. In the same way, parents’ social-status affects social compromise (parameter $\gamma_1$) significantly, no other than the entire sample. In addition, psychological oppression within family has a positive and significant direct effect on social compromise (parameter $\gamma_2$). Considering only CASE II, stress development due to the lack of learning skills influences non Australian-born student’s social compromise positively and significantly (parameter $\gamma_3$).

The estimated total effects of parents encouraging environment on creativity and social compromise is significant (allowing for both cases, I and II),
TABLE 1. Estimates of the Model (ML)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Entire sample</th>
<th>Sub-sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CASE I</td>
<td>CASE II</td>
</tr>
<tr>
<td>( \beta )</td>
<td>-0.87*</td>
<td>-22.82</td>
</tr>
<tr>
<td>( \gamma_1 )</td>
<td>-0.08***</td>
<td>-1.64</td>
</tr>
<tr>
<td>( \gamma_2 )</td>
<td>0.86*</td>
<td>15.37</td>
</tr>
<tr>
<td>( \gamma_3 )</td>
<td>0.05</td>
<td>1.22</td>
</tr>
<tr>
<td>( \lambda_1 )</td>
<td>2.01*</td>
<td>11.95</td>
</tr>
<tr>
<td>MFF ( \chi^2 )</td>
<td>10.84</td>
<td>6.31</td>
</tr>
<tr>
<td>NTWLS ( \chi^2 )</td>
<td>10.95</td>
<td>6.22</td>
</tr>
</tbody>
</table>

*\( p < .01; **p < .05; ***p < .10.\)

TABLE 2. Estimated Total Effects of Ksi on Eta

<table>
<thead>
<tr>
<th>Entire sample</th>
<th>Sub-sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>CASE I</td>
<td>CASE II</td>
</tr>
<tr>
<td>PARSTAT</td>
<td>PARREP</td>
</tr>
<tr>
<td>CREAT*</td>
<td>- .07***</td>
</tr>
<tr>
<td>(1.64)</td>
<td>(-10.34)</td>
</tr>
<tr>
<td>COMPR*</td>
<td>- .08***</td>
</tr>
<tr>
<td>(-1.64)</td>
<td>(15.37)</td>
</tr>
</tbody>
</table>

*\( p < .01; **p < .05; ***p < .10.\)

TABLE 3. Estimated Total Effects of Eta on Eta

<table>
<thead>
<tr>
<th>Entire sample</th>
<th>CASE I</th>
<th>Sub-sample</th>
<th>CASE II</th>
</tr>
</thead>
<tbody>
<tr>
<td>CREAT*</td>
<td>-</td>
<td>-.87*</td>
<td>-</td>
</tr>
<tr>
<td>(19.82)</td>
<td>(-19.54)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COMPR*</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*\( p < .01.\)

and parents’ social-status affects the estimated total effect of student’s creativity and social compromise significantly, nothing but CASE I. At the same time, the estimated total relationship of social compromise on creativity is negative and highly significant. The implication of these estimates will be discussed in the next section.
5 Concluding remarks

The most important results of this paper can be reviewed as follows:

Considering CASE II, and I the estimates show a negative and statistically significant relationship between creativity and social compromise.

In view of CASE I: (1) Social compromise is directly and positively influenced by oppression within family and negatively influenced by parents’ social-status. (2) Creativity is indirectly and negatively influenced by psychological oppression within family and positively influenced by parents’ social-status.

In consideration of CASE II, the estimates demonstrate a direct and positive influence of oppression within family and student’s stress on social compromise, and an indirect and negative effect of them on creativity.

These findings may be delved into by a follow-up survey to explore the effects of the Australian educational system on the working lives of Australia’s young generations. The non Australian-born students, coming from a different schooling environment on the contrary to their Australia-born classmates, have to struggle to acquire some learning skills.

All parents, regardless of their ethnicity, wish and “push” their children to get a university degree, which is considered to be a prerequisite for a prestigious and wealthy career. According to our findings, Australian parents, having higher average educational level than that of non Australians’, seem to be interested in being their children qualified rather for a prosperous than for a high-status life; there is a hint from our data that non Australian parents consider a university degree as the “passport” to a high-status life, which in consequence is an easy way out from their social-isolation. Parents interfere with their children’s life choices at the expense of creativity, and do not become conscious that social-compromise is responsible for people’s unhappiness and thus a stifling of one’s creativity.

Australian university students may be suffering a serious strangling of their creativity because of a prevailing repressive attitude within Australian and non-Australian families and student’s stress development due to the absence of necessary learning skills only among non Australian-born students. Thus, according to our findings, the student’s creativity reinforcement passes through two different channels: that one of family and the other one of education.

References


LISREL Website (2002). *Historical Background,* www.ssicentral.com/lisrel/mainlis.htm
Linear mixed-effects modelling of panel data:
a cautionary note

Iraj Kazemi

1 Department of Statistics, Isfahan University, Isfahan, Iran.

Abstract: This paper addresses various pitfalls that can typically occur in the
application of linear mixed-effects modelling to panel data. A special focus is
made to highlight some new topics that deserve more attention. Specifically, a
modified version of the \( h \)-likelihood approach is introduced to handle the problem
of negative variance estimates. An extension is also proposed to deal with the
initial conditions of the dynamic process. Evidence from Monte Carlo simulations
is provided to show the usefulness of the extensions. An empirical example is
presented to a real data set.

Keywords: \( h \)-likelihood; panel data; the initial conditions; the negative estimate
of variance components.

1 Introduction

Mixed-effects models are useful for modelling the unobserved individual
heterogeneity effects frequently present in panel data. They allow for the
responses to be correlated over time through the individual effects presented
in the mean structure. However, in order to draw meaningful inference from
such data, model specification and related estimation issues have to be ad-
dressed carefully. In such models, the likelihood of all unknown quantities
is based on the joint density of responses and the individual effects. This
extended likelihood called hierarchical (\( h \)-) likelihood by Lee and Nelder
(1996). Two specific issues then come up with interesting recommendations
in using the \( h \)-likelihood. The first important issue is credibly accounting
for the effect of unobserved initial conditions of the dynamic process (An-
derson and Hsiao, 1982). The second issue relates to the occurrence of a
negative estimate for the variance of the individual effects. A vast ma-
jority of the published papers do not address the negativity problem at
all. This paper proposes strategies for coping with these problems that are
rarely addressed in the literature. Furthermore, the paper emphasizes that
the negativity problem may occur because of model misspecification due
to the initial conditions.
2 Model Specification

In panel studies, the components of an observation \( \mathbf{Y}_i = (Y_{i1}, \cdots, Y_{iT})' \) represent a sequence of responses at times 1, \( \cdots, T \) for individual \( i = 1, \cdots, N \). Suppose that \( \mathbf{Y}_{i,t-1} \) is a \( T \)-element vector of one-period lagged values of \( Y \), and \( \mathbf{X}_i \) is a \( T \times K \) matrix of explanatory variables. The most commonly used linear mixed-effects model for \( \mathbf{Y}_i \), as described by Laird and Ware (1982), is

\[
\mathbf{Y}_i = \tilde{\mathbf{X}}_i \theta + \mathbf{Z}_i \alpha_i + \varepsilon_i, 
\]

where \( \tilde{\mathbf{X}}_i = (\mathbf{Y}_{i,t-1} \mathbf{e}_T \mathbf{X}_i) \), \( \mathbf{e}_T \) is a vector of ones of order \( T \), \( \theta = (\gamma \lambda \beta)' \), and \( \mathbf{Z}_i \) equals to \( \mathbf{e}_T \). The usual assumptions are that \( \alpha_i \sim \text{iid} \mathcal{N}(0, \sigma^2_\alpha) \), \( \varepsilon_i \sim \text{iid} \mathcal{N}(0, \sigma^2_\varepsilon \mathbf{I}_T) \), where \( \mathbf{I}_T \) denotes an identity matrix, the \( \alpha_i \) and the \( \varepsilon_i \) are independent, and the \( \mathbf{X}_{it} \) are strictly exogenous with respect to \( \alpha_i \) and \( \varepsilon_i \).

3 Some Modelling Issues with Mixed Effects

3.1 The Initial Conditions

Although mixed models offer great flexibility for modelling the unobserved individual heterogeneity effects, they suffer practically from some fundamental problems. One important issue relates to the initial conditions of the dynamic process. It follows from the conditional distributions of \( Y_{it} \) and \( \alpha_i \) given \( Y_{i,t-1} \), for \( t \geq 1 \), that the standard \( h \)-likelihood is of the form

\[
L(\theta, \sigma, \alpha) = \prod_{i,t} f(y_{it}|y_{i,t-1}, \cdots, y_{i0}; \alpha_i, \theta, \sigma^2_\varepsilon) \times \prod_i f(\alpha_i | \sigma^2_\alpha). \quad (2)
\]

The first term in (2) corresponds to the likelihood function of so-called fixed-effects models in panel data studies, while the second product corresponds to the density function of the effects \( \alpha_i \). In order to examine the modelling issues associated with the initial conditions problem, let us define the full \( h \)-likelihood function based on the joint density of the initial and subsequent outcomes with the individual effects, as follows

\[
L(\theta, \sigma^2, \phi, \alpha) = \prod_{i,t} f(y_{it}|y_{i,t-1}, \cdots, y_{i0}, \alpha_i; \theta, \sigma^2_\varepsilon) \times \prod_i f(y_{i0}, \alpha_i | \phi), \quad (3)
\]

where \( \phi \) is a vector of model parameters specified in the joint density \( f(y_{i0}, \alpha_i) \). The classical \( h \)-likelihood approach assumes that the initial outcomes at time 0 are uncorrelated with the effects \( \alpha_i \), i.e., \( f(y_{i0}, \alpha_i) = f(y_{i0}) f(\alpha_i) \). Operating this restriction, the full \( h \)-likelihood (3) then reduces to a product of two factors: the first part corresponds to the classical \( h \)-likelihood (2), and the second to a marginal likelihood corresponding to the density of \( Y_{i0} \) for all \( i \). This leads to the initial conditions problem because, in general, the effects \( \alpha_i \) are correlated with the initial state \( Y_{i0} \).
3.2 A Negative Estimate of Variance

The principal interest in mixed-effects modelling is focused on the estimation of both regression coefficients, \( \theta \), and the variance components, \( \sigma_\epsilon^2 \) and \( \sigma_\alpha^2 \), as well as the individual effects, \( \alpha_i \), using the \( h \)-likelihood approach. For given values of \( \sigma' = (\sigma_\epsilon^2, \sigma_\alpha^2) \), the parameter \( \theta \) and the effects \( \alpha_i \) are estimated jointly by maximizing the logarithm of (2). Moreover, Lee and Nelder (1996) use a modified profile likelihood to estimate the variance components. Specifically, using the equation (1) and simplifying the corresponding expressions,

\[
\hat{\theta} = (\psi B_{\bar{\alpha} \bar{\epsilon}} + W_{\bar{\alpha} \bar{\epsilon}})^{-1} (\psi b_{\bar{\alpha}y} + w_{\bar{\alpha}y}),
\]  

and

\[
\hat{\alpha}_i = \frac{T}{T + \kappa} \bar{r}_i, \quad \text{for } i = 1, \ldots, N,
\]  

where \( B \) and \( W \) refer to the between- and within-individual variations, \( \psi = \sigma_\epsilon^2 / (T \sigma_\alpha^2 + \sigma_\epsilon^2) \), the \( \bar{r}_i \)'s denote the average of residuals \( r_{it} = y_{it} - \hat{X}_{it}'\hat{\theta} \) over time periods for individual \( i \), and \( \kappa = \sigma_\epsilon^2 / \sigma_\alpha^2 \). Furthermore,

\[
\hat{\sigma}_\epsilon^2 = \frac{\mathbf{e}'\mathbf{e}}{NT - df},
\]  

and

\[
\hat{\sigma}_\alpha^2 = \frac{1}{N} \sum_i \hat{\alpha}_i^2 + \frac{\hat{\sigma}_\epsilon^2}{T + \kappa},
\]  

where \( \mathbf{e} \) is the vector of model residuals \( e_{it} = y_{it} - \hat{X}_{it}'\hat{\theta} - \hat{\alpha}_i \), and \( df = N (1 + \kappa / T)^{-1} \). The maximum \( h \)-likelihood estimates are computed by iterating between these estimating equations until convergence.

A serious disadvantage of the \( h \)-likelihood approach is that it may produce a negative estimate of \( \sigma_\alpha^2 \), as is well known in the variance components literature. The implementation of an inequality constraint in the maximization process gives also a zero estimate which points to refitting the model using the conditional likelihood \( L(\theta, \sigma_\epsilon^2 | \alpha, \mathbf{y}) \). To illustrate the problem, after some algebra has been done, the estimate \( \hat{\sigma}_\alpha^2 \) may be simplified as \( \hat{\sigma}_\alpha^2 = (\hat{\sigma}_\epsilon^2 - \hat{\sigma}_\alpha^2) / T \), where \( \hat{\sigma}_\epsilon^2 = (T / N) \sum_i \bar{r}_i^2 \) is the between-individual mean squares residual, and \( \hat{\sigma}_\epsilon^2 \) is the residual variance estimate representing the within-individual variation. Thus, if \( \hat{\sigma}_\epsilon^2 \) is substantially less than \( \hat{\sigma}_\alpha^2 \), there is no guarantee that the solution of mixed equations for \( \sigma_\alpha^2 \) be non-negative. This result makes a challenge in introducing non-negative variance estimates using suitable alternative estimation approaches. Some ways of coping with this problem and related issues are discussed in the forthcoming section.
4 Alternative Estimation Approaches

4.1 A Modified H-Likelihood

Depending on the application of mixed-effects models, various modifications of the \( h \)-likelihood have been suggested (Lee and Nelder, 2001) to address particular issues. We propose here an alternative approach to deal with the negativity problem. The idea is motivated, from a Bayesian point of view, by treating the variance components as random variables whose distributions belong to some parametric family of distributions that support non-negative values. This alternative can be motivated as follows.

The aim is to obtain a nonzero value for the variance component \( \sigma^2_\alpha \) when the standard \( h \)-likelihood (2) has a maximum at \( \sigma^2_\alpha = 0 \). Assuming that \( \sigma^2_\varepsilon \) and \( \sigma^2_\alpha \) are independent with density functions \( \pi(\sigma^2_\varepsilon) \) and \( \pi(\sigma^2_\alpha) \), respectively, the modification is made by multiplying the \( h \)-likelihood function (2) in these densities. It gives

\[
L_m(\theta, \sigma, \alpha) = L(\theta, \sigma^2_\varepsilon | \alpha, y) \ f(\alpha | \sigma^2_\alpha) \ \pi(\sigma^2_\varepsilon) \ \pi(\sigma^2_\alpha),
\]

which represents a likelihood based on the joint distribution of \( Y, \alpha, \) and \( \sigma \). This consists of the standard \( h \)-likelihood (2) conditional on variance parameters \( (\sigma^2_\varepsilon, \sigma^2_\alpha) \), plus an additional factor for the joint density function of those variances. The effect of this additional factor, as will be shown, is that \( \sigma^2_\alpha \) is no longer estimated negative or even zero.

To operationalise the estimation process using (8), it is required to specify distribution functions for the variance components. Various families of distributions have been suggested (e.g. Browne and Draper, 2000). A commonly useful family is provided by the inverse-gamma distributions. That is, it is assumed that \( \sigma^2_\varepsilon \sim IG \left( m_2, \delta^2 \right) \) and \( \sigma^2_\alpha \sim IG \left( n_2, \upsilon^2 \right) \), where \( m, n, \delta, \) and \( \upsilon \) are pre-specified positive scalars. Given this specification, the maximization of (8) straightforwardly shows that the estimates for \( \theta \) and \( \alpha_i \) are given by (4) and (5). The variance components can also be estimated using a modified profile likelihood when the term \( \log L(\theta, \sigma, \alpha) \) is substituted by \( \log L_m(\theta, \sigma, \alpha) \). Specifically, it can be shown that

\[
\hat{\sigma}^2_{\varepsilon,m} = \frac{(T - 1) \hat{\sigma}^2_\varepsilon + \left( \frac{\hat{\kappa}_m}{T + \hat{\kappa}_m} \right)^2 \hat{\sigma}^2_\varepsilon + \frac{4}{N}}{T - 1 + \frac{\hat{\kappa}_m}{T + \hat{\kappa}_m} + \frac{m + 2}{N}},
\]

and

\[
\hat{\sigma}^2_{\alpha,m} = \frac{(1 - \frac{1}{T + \hat{\kappa}_m}) \hat{\sigma}^2_\alpha + \frac{\upsilon}{NT}}{1 + \frac{1}{T + \hat{\kappa}_m} + \frac{n + 2}{NT}},
\]

where \( \hat{\sigma}^2_\varepsilon \) and \( \hat{\sigma}^2_\alpha \) are defined previously. This approach performs well as it always yields non-negative estimates of \( \sigma^2_\alpha \). The success of this estimation
technique, however, depends on the suitable choices of \( m, n, \delta, \) and \( \nu. \) In practice very small positive numbers are chosen to ensure that the \( h \)-likelihood is not dominated by density functions \( \pi (\sigma_{2}^{2}) \) and \( \pi (\sigma_{2}^{2}) \), and that the modified \( h \)-likelihood is driven by the data.

### 4.2 The Full H-Likelihood

The modified \( h \)-likelihood approach is generally regarded as superior to classical maximum \( h \)-likelihood. However, it offers no rule for handling the initial conditions problem, implying that the estimation results are to be treated with caution. Thus, the second strategy for handling both the negativity and the initial conditions problems is motivated by the use of full \( h \)-likelihood (3) which essentially models the joint distribution of \( Y_{i0} \) and \( \alpha_{i} \). Specifically, it can be shown that the maximization of the full \( h \)-likelihood (2) with respect to the parameters that are within the parameter space results in a non-negative estimate of variance, if there exists a nonzero correlation between \( Y_{i0} \) and \( \alpha_{i} \) (Kazemi and Crouchley, 2006).

### 5 Simulation Studies

Two Monte Carlo simulations are performed to address the modelling issues. The response variable for the subsequent outcomes is generated according to

\[
Y_{it} = \lambda + \gamma Y_{i,t-1} + \alpha_{i} + \varepsilon_{it},
\]

for \( N=100 \) and \( T=5, \) where the \( Y_{i0} \sim iidN (\lambda_0, \sigma_0^2) \). In the first study, we set \( \gamma = 0.6, \lambda = -0.5, \lambda_0 = -1, \sigma_0^2 = \sigma_0^2 = 1.0, \) and \( \text{cov} (Y_{i0}, \alpha_{i}) = \sigma_{0\alpha} = 0.5. \) Simulation proceeds by sequentially generating the \( \varepsilon_{it} \) from a standard normal distribution and independent pairs of values \( (Y_{i0}, \alpha_{i}) \) from a bivariate normal distribution with the above specification.

<table>
<thead>
<tr>
<th>Table 1. Estimation results for simulated data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>( \gamma )</td>
</tr>
<tr>
<td>( \lambda )</td>
</tr>
<tr>
<td>( \lambda_0 )</td>
</tr>
<tr>
<td>( \sigma_0^2 )</td>
</tr>
<tr>
<td>( \sigma_{0\alpha} )</td>
</tr>
</tbody>
</table>

Standard errors are in parantheses.
Results are presented in Table 1. By treating $Y_{i,t-1}$ conventionally as an explanatory variable the model can be easily implemented in a standard software, such as the SAS procedure MIXED. It is seen that this treatment gives rise to an upwardly biased estimate of the true $\gamma$ and to under-estimation of $\sigma^2_\alpha$. The relative bias for $\hat{\gamma}$ is more than 10 percent, showing that the magnitude of bias is considerable. It can be further shown that the bias in $\hat{\gamma}$ increases with the correlation between $Y_{i0}$ and $\alpha_i$.

In order to operationalise the modified $h$-likelihood approach, the values for $m$, $n$, $\delta$, and $\nu$ are first assumed all equal to the small value $\epsilon$. Then, we fit the model with a range of values for $\epsilon$ to check the sensitivity of the results. It was found that results are generally insensitive to the specific choice of $\epsilon$ in the region of 0.001, so I report findings with this value. Results show that the modified version also produces biased results. In contrast, the estimated values using the full $h$-likelihood approach come very close to the true values, implying that this approach is useful in devising unbiased estimates.

Table 2. Estimation results: a negative estimate of variance

<table>
<thead>
<tr>
<th>Parameter restriction for $\sigma^2_\alpha$</th>
<th>Parameter</th>
<th>The classical $h$-likelihood</th>
<th>The classical $h$-likelihood</th>
<th>The modified version</th>
<th>The full $h$-likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2_\alpha \in \mathbb{R}$</td>
<td>$\beta$</td>
<td>0.6425 (0.0070)</td>
<td>0.6388 (0.0073)</td>
<td>0.6307 (0.0081)</td>
<td>0.6005 (0.0115)</td>
</tr>
<tr>
<td></td>
<td>$\lambda$</td>
<td>-0.1835 (0.0079)</td>
<td>-0.1870 (0.0083)</td>
<td>-0.1946 (0.0094)</td>
<td>-0.2231 (0.0122)</td>
</tr>
<tr>
<td></td>
<td>$\gamma_0$</td>
<td>-1.3898 (0.1002)</td>
<td>-1.3898 (0.1002)</td>
<td>-1.3898 (0.1002)</td>
<td>-1.3898 (0.1002)</td>
</tr>
<tr>
<td></td>
<td>$\sigma^2_{\epsilon}</td>
<td>0.0116 (0.0008)</td>
<td>0.0112 (0.0007)</td>
<td>0.0107 (0.0007)</td>
<td>0.0107 (0.0007)</td>
</tr>
<tr>
<td></td>
<td>$\sigma^2_\nu$</td>
<td>-0.0004 (0.0003)</td>
<td>0.0000 (0.0000)</td>
<td>0.0004 (0.0004)</td>
<td>0.0004 (0.0004)</td>
</tr>
<tr>
<td></td>
<td>$\sigma_{\nu \alpha}$</td>
<td>0.0112 (0.0007)</td>
<td>0.0107 (0.0007)</td>
<td>0.0101 (0.0005)</td>
<td>0.0101 (0.0005)</td>
</tr>
<tr>
<td></td>
<td>$\sigma_{\nu \beta}$</td>
<td>0.0100 (0.0005)</td>
<td>0.0100 (0.0005)</td>
<td>0.0100 (0.0005)</td>
<td>0.0100 (0.0005)</td>
</tr>
<tr>
<td></td>
<td>$\sigma_{\nu \gamma}$</td>
<td>0.0316 (0.0086)</td>
<td>0.0316 (0.0086)</td>
<td>0.0316 (0.0086)</td>
<td>0.0316 (0.0086)</td>
</tr>
<tr>
<td>Obs.</td>
<td>500</td>
<td>500</td>
<td>500</td>
<td>600</td>
<td>600</td>
</tr>
</tbody>
</table>

Standard errors are in parantheses.

To address the negativity problem, $\sigma^2_\alpha$ is considered to be small relative to $\sigma^2_\beta$. Specifically, we set $\sigma^2_\delta = 0.001$, $\sigma^2_\epsilon = 0.01$, and $\sigma_{\alpha} = 0.03$. Results are presented in Table 2. The model is fitted in two ways. First, the SAS procedure MIXED is used but without specifying any restrictions on bounds for the variances. Although the procedure MIXED reports the convergence criteria met, it fails to produce a non-negative estimate of variance. In the second case, the model is refitted under the restriction that the variances take non-negative values. The parameter $\sigma^2_\alpha$ is now estimated to be zero. The modified $h$-likelihood gives a positive estimate and when compared with its standard error, suggests significant variability among the individuals. This approach, however, completely ignores the covariance structure of the true model, resulting in biased results. The
evidence in Table 2 illustrates that the full $h$-likelihood is superior among these approaches even for such a defective model. Using a non-negative optimization method, parameters are estimated close to the true values. Further, the covariance $\sigma_{\alpha \alpha}$ is significant, emphasizing that the problem of negative estimates may be related to the model misspecification due to the incorrect treatment of the initial conditions.

6 Empirical Analysis

For the purpose of illustration, an empirical examples is presented. The example is an economic growth model which is taken from classical dynamic growth studies (e.g. Nerlove, 1999) based on annual observations for 94 countries over the period 1960-85. The model is derived from assumptions about a production function and inclusion in the specification of the savings rate, $s$, the population growth rate, $n$, the rate of technical progress, $g$; and the depreciation rate, $\delta$. Suppose $GDP_{it}$ to be the logarithm of per capita gross domestic product for country $i$ at time $t$, the model is given by

$$GDP_{it} = \lambda + \gamma GDP_{i,t-1} + \beta X_{it} + \alpha_i + \varepsilon_{it}, \ t = 1965, 1970, 1975, 1980, 1985,$$

(12)

where $X_{it} = \log(s_{it}) - \log(n_{it} + g + \delta)$, $\alpha_i$ is a country-specific effect, and $\varepsilon_{it}$ is the error term. As in the empirical growth models, it is assumed that $g$ and $\delta$, summing to 0.05, are constant across countries. In fitting model (12), it is argued that the $\alpha_i$ representing technological differences between countries are correlated with the initial level of income which necessitates modelling of the the initial conditions carefully.

Table 3. Estimation results for different estimation approaches.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>The classical $h$-likelihood</th>
<th>The modified version</th>
<th>The full $h$-likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>0.6987 (0.0789)</td>
<td>0.6987 (0.0789)</td>
<td>0.6424 (0.0757)</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>7.0635 (0.0757)</td>
<td>7.0635 (0.0757)</td>
<td>7.0872 (0.0753)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.9339 (0.0122)</td>
<td>0.9327 (0.0123)</td>
<td>0.8393 (0.0207)</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.1370 (0.0131)</td>
<td>0.1379 (0.0132)</td>
<td>0.1957 (0.0147)</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>0.5162 (0.0874)</td>
<td>0.5244 (0.0880)</td>
<td>1.2002 (0.1523)</td>
</tr>
<tr>
<td>$\sigma_\varepsilon$</td>
<td>0.4353 (0.0635)</td>
<td>0.4353 (0.0635)</td>
<td>0.4377 (0.0643)</td>
</tr>
<tr>
<td>$\sigma_{\alpha \alpha}$</td>
<td>.</td>
<td>.</td>
<td>0.0576 (0.0134)</td>
</tr>
<tr>
<td>$\sigma_{\beta \beta}$</td>
<td>0.0172 (0.0013)</td>
<td>0.0170 (0.0013)</td>
<td>0.0153 (0.0011)</td>
</tr>
<tr>
<td>$\sigma_{\gamma \gamma}$</td>
<td>0.0022 (0.0010)</td>
<td>0.0022 (0.0010)</td>
<td>0.0103 (0.0034)</td>
</tr>
</tbody>
</table>

Obs 564 564 564
Standard errors are in parentheses.

AIC -325.278 -316.961 -365.3059
To do this, a realistic way is to add a flexible reduced form equation

\[ GDP_i = \lambda_0 + X_i \beta_0 + u_{i0}, \quad i = 1, \cdots, 94, \quad (13) \]

where the \( X_i \)'s are supposed to be uncorrelated with \( u_{i0} \), the \( u_{i0} \) and \( \varepsilon_{it} \) to be uncorrelated for all \( t > 0 \), the initial error \( u_{i0} \sim \text{iid} N (0, \sigma_0^2) \), and \( \text{cov}(u_{i0}, \alpha_i) = \sigma_{0\alpha} \). With these assumptions, as time goes on, \( \alpha_i \) affects \( Y_{it} \) in all subsequent periods, including \( Y_{i0} \). This pragmatic solution for the \( Y_{i0} \) is proposed by Kazemi and Crouchley (2006) who argue that the approach performs with greater flexibility than other start-up models.

To compare estimation approaches, the model is fitted using the \( h \)-likelihood (2) and the modified version where the \( Y_{i0} \) are assumed to be uncorrelated with \( \alpha_i \). Results are shown in Table 3, together with those obtained by using the full \( h \)-likelihood (3) which includes the information of the 1960 GDP per capita taken into account. Comparing the estimates for \( \gamma \), the classical \( h \)-likelihood and its modified version overestimates the true \( \gamma \). Most importantly, results suggest that it is questionable to assume the dependency term \( \sigma_{0\alpha} \), representing the association between the initial observations and the subsequent process, is zero. Specifically, the \( AIC \) values indicate the unconditional model is more appropriate given the data.

References


Survival models for recurrent events in the estimation of the effect of badger culling on tuberculosis in Irish cattle herds

G. E. Kelly\textsuperscript{1}, J. Condon\textsuperscript{2}

\textsuperscript{1} UCD School of Mathematical Sciences, University College Dublin, Ireland
\textsuperscript{2} Centre for Statistical Science and Operational Research, Queen’s University Belfast, Northern Ireland.

Abstract: Survival models for recurrent events are considered. In particular we consider frailty models in conjunction with the Cox model. Klein(1992) derived an EM algorithm to estimate the parameters in the model assuming gamma frailty. We adapt his algorithm to nonparametric frailty distributions. The methods developed are applied to the East Offaly Project data to examine the effect of badger culling on tuberculosis in cattle herds, adjusting for potentially correlated recurrent events. The algorithms are further explored by considering two examples from the literature and presenting some simulation results.

Keywords: Censored survival data; proportional hazards regression; gamma frailty; nonparametric frailty; EM algorithm.

1 Introduction

In this paper we examine the issues that arise in modelling recurrent events in survival analysis. Our primary example consists of times of trade restriction due to TB of herds in the area of East Offaly in Ireland. In part of the area, the treatment area, badgers were proactively removed over some years. In the control area minimal badger removal occurred and then only reactively. We consider the time period 1989-2004. Further details of the study - the East Offaly Project (EOP) may be found in O’ Mairtin et al. (1998).

Several approaches to modelling such survival data (i.e. times to restriction) have been outlined in the literature. Therneau and Grambsch (2000) outline both a marginal models approach and a random effects or frailty approach. We consider both. For the marginal model we consider the Anderson-Gill (AG) approach to modelling multiple events which assumes mutual independence of the observations within a herd. However the standard variance estimates are replaced by ones which are corrected for the possible correlations using a grouped jackknife.

For the frailty approach parametric forms for the frailty distribution - the
272 Survival models for recurrent events

gamma and log-normal are fitted. We also consider a nonparametric distribution for the frailty and compare the results.

We develop two algorithms for fitting the nonparametric frailty distribution. In this we follow the work of Hinde (1982) and Klein (1992). The algorithms are applied to a further two examples from the literature. In addition we present some preliminary simulation results.

2 Frailty models

We have data consisting of time to trade restriction or time on study and covariate values of n herds from a population. A herd may have more than one trade restriction over time leading to multiple survival times. Denote the time of the jth event or the end of study, whichever occurs first, for the ith herd by \( t_{ij} \), \( j = 1, \ldots, n \); \( i = 1, \ldots, n \). We denote the associated censoring information by \( \delta_{ij} \) (\( \delta_{ij} = 1 \) if the \( t_{ij} \) is a trade restriction; \( \delta_{ij} = 0 \) otherwise). Let \( x_i(t) = (x_{i1}(t), \ldots, x_{ip}(t)) \) denote the possibly time dependent covariates associated with the ith herd. The one or more observations on the same herd constitute a 'family'. This 'family' consists of dependent event times due to some unobserved covariate information summarized in a frailty, \( \omega_i \). Note that families of size 1 are allowed. It is assumed that the conditional on the frailty \( \omega_i \), the hazard rate for the jth restriction on the ith herd is:

\[
\lambda(t/x_i(t), \omega_i) = \omega_i \lambda_0(t) e^{x_i \beta} = \omega_i \lambda_i(t).
\]

The integrated hazard will be denoted by \( \omega_i \Lambda_i(t) \) where

\[
\Lambda_i(t) = \int_0^t Y_i(u) \lambda_0(u) e^{x_i(u) \beta} du,
\]

where \( Y_i(u) \) is an observation process for herd i, defined as \( Y_i(u) = 1 \) if herd i is at risk at time \( u \) and zero otherwise. Klein (1992) considered the case where the \( \omega_i \)'s were independent and identically distributed gamma variates and extended the Cox regression model accordingly. We consider a nonparametric distributions for the frailty. We assume the unobserved frailties \( \omega_i, i = 1, \ldots, n \) come from a discrete distribution with mass points \( h_1, \ldots, h_k \) and associated probabilities \( \pi_1, \ldots, \pi_k \), i.e.

\[
f(\omega) = \Pi_{l=1}^k \pi_l I_{h_l}(\omega)
\]

The joint log likelihood for the data and unobserved frailties is

\[
log(L_{full}) = \sum_i D_i ln(\omega_i) + \sum_i \sum_j \delta_{ij} ln(\lambda_i(t_{ij})) - \sum_i \omega_i \Lambda_i(t_{in}) + \sum_{i=1}^n ln f(\omega_i) = L_1 + L_2
\]

where

\[
L_1 = \sum_i D_i ln(\omega_i) + \sum_i ln f(\omega_i)
\]
and

\[ L_2 = \sum_i \left[ \sum_j \delta_{ij} \ln \lambda_i(t_{ij}) \right] - \omega_i \lambda_i(t_{in_i}) \]  

(4)

and \( D_i = \sum_{j=1}^{n_i} \delta_{ij} \) is the observed number of events for herd \( i \). Let \( d_{ij} \) be the number of tied events at \( t_{ij} \). Klein (1992) showed that the M step of the E-M algorithm involves \( L_1 \) and \( L_2 \) separately. He showed, given current estimates \( \hat{\omega}_i \), maximizing \( E(L_2/\beta, \lambda_0) \) with respect to the unknown parameter \( \beta \) can be accomplished by setting (with a slight modification for the AG model)

\[ \hat{\lambda}_0(t_{ij}) = \frac{d_{ij}}{\sum_{m \in R(t_{ij})} \omega_m e^{x'_m(t_{ij}) \beta}}. \]  

(5)

and substituting in to yield the profile log-likelihood

\[ L_3(\beta) = \sum_i \left[ \sum_j \delta_{ij} (\ln \hat{\lambda}_0(t_{ij}) + x'_i(t_{ij}) \beta) \right] + \sum_i (\hat{\omega}_i \sum_{t_{in_i} \leq t_{ij}} e^{x'_i(t_{ij}) \beta} \hat{\lambda}_0(t_{ij}) Y_i(t_{ij})) \]  

(7)

\[ = \sum_i \left[ \sum_j \delta_{ij} \left( -\ln \sum_{m \in R(t_{ij})} \omega_m e^{x'_m(t_{ij}) \beta} \right) + x'_i(t_{ij}) \beta \right] + \sum_i D_i \]  

(8)

\( L_3(\beta) \) is then maximized. We note that this log partial likelihood is of the form of a usual partial likelihood for the Cox model with the inclusion of an additional covariate \( \ln[\omega_m] \) with known coefficient of 1. Therefore, a standard Cox regression program can be modified to obtain the updated estimate of \( \beta \). To maximize \( E(L_1) \) we proceed as follows. The marginal likelihood for the \( i \)th family is

\[ L_i = E(L_{full_i}) = \sum_{l=1}^{h_{max}} \left( \prod_{j=1}^{n_{max}} [h_l \lambda_l(t_{ij})]^{\delta_{ij}} e^{-h_l \lambda_l(t_{in_i}) \pi_l} \right) \]  

(9)

Thus the conditional density of \( \omega_i \) given the data is \( L_{full_i}/L_i \) has mass points \( h_1, \ldots, h_k \) and the probability associated with \( h_l \) given by

\[ \pi_{il *} = \frac{h_l^{D_l} e^{-h_l \lambda_l(t_{in_i}) \pi_l}}{\sum_{r=1}^{k} h_r^{D_r} e^{-h_r \lambda_r(t_{in_i}) \pi_r}} \]  

(10)

The mean of \( \omega_m \) with respect to this conditional density is

\[ \bar{\omega}_m = E(\omega_m) = \sum_{l=1}^{k} h_l \pi_{ml *}. \]  

(11)

Thus

\[ E(L_1) = \sum_i D_i E(\ln(\omega_i)) + \sum_i E(\ln f(\omega_i)) \]

\[ = \sum_{i=1}^{n} D_i \sum_{l=1}^{k} [\ln(h_l) \pi_{il *}] + \sum_{i=1}^{n} \sum_{l=1}^{k} (\ln(\pi_{il})) \pi_{il *}. \]  

(12)

The estimating algorithm proceeds by first making an initial guess at the values, so we start with \( h_1, \ldots, h_k, \pi_1, \ldots, \pi_k \), subject to the constraint
that $\Sigma_l h_l \pi_l = 1$ i.e. start with $\hat{\omega}_m = 1$. As in Klein the EM algorithm proceeds as follows:

**Algorithm 1**

**Step 1** Using a standard Cox regression program, obtain initial estimates of $\beta$, $\lambda_0$ and $C_i$ from (8), (5) and (6) respectively, with $\hat{\omega}_m = 1$.

**Step 2** Using the current values of $\hat{h}_l$, $\hat{\pi}_l$, $l = 1, \ldots, k$ and $C_i$ compute $\pi_{ml}^*$ and $\hat{\omega}_m$ from (10) and (11).

**Step 3** Update the estimate of $\hat{\pi}_l$ and $\hat{h}_l$ using (12) (regarding (12) as a function of the $\pi_l$ and $h_l$), subject to the constraints that $\Sigma_l \pi_l = 1$ and $\Sigma_l h_l \pi_l = 1$. Update the estimate of $\beta$ (and $\lambda_0$, $C_i$) using (8), (5) and (6).

**Step 4** Iterate between steps 2 and 3 until convergence. A further refinement of Step 3 is possible. Differentiating equation (12) with respect to the $\pi_l$ yields

$$\pi_l = \frac{\Sigma_i \pi_{il}^*}{n}$$

Thus, this can be substituted into equation (12) to get a function of the $h_l$ only and iterate between equation (12) and (13) before continuing to Step 2.

A second algorithm motivated by Hinde (1982) and Hartzel et al. (2001) can be formulated by iterating between the equations in a different way. The joint log likelihood is split differently by defining $L_1$ and $L_2$ by

$$L_1 = \Sigma_i \ln f(\omega_i)$$

and

$$L_2 = \Sigma_i [\Sigma_j \delta_{ij} (\ln \omega_i + \ln \lambda_i(t_{ij}))] - \omega_i \Lambda_i(t_{im})$$

Equation (11) for $\hat{\omega}_m$ is substituted into equations (14) to give

$$L_3 = E[L_1] = \frac{\Sigma_i \sum_{l=1}^k (\ln \pi_l) \pi_{il}^*}{n}$$

and also substituted into equation (15) after performing the same profile likelihood step to that leading to equation (8) to give

$$L_4 = E[L_2] = \Sigma_i \Sigma_j \delta_{ij} \left(-\ln [\Sigma_{j \in R(t_{ij})} (\Sigma_{l=1}^k h_l \pi_{ml}^*) e^{x_{il}(t_{ij})\beta}] + \sum_{l=1}^k \ln (h_l) \pi_{ml}^* + x_{il}(t_{ij})\beta \right) \Lambda_i(t_{im}) + \Sigma_i D_i$$

$L_3$ and $L_4$ are maximized separately. Maximizing $L_3$ gives $\pi_l$ as in (13). We note that $L_4$ is no longer of the form of a partial likelihood for a Cox model or even a weighted partial likelihood. However it can be maximized using a nonlinear maximization routine. The second algorithm proceeds as
follows.

**Algorithm 2**

**Step 1** Start with initial weights $\pi_{il^*}, i = 1, \ldots, n; l = 1 \ldots, k$ (e.g. giving them equal value such that for each $i$ $\pi_{il^*} = 1/k$). Compute $\pi_l$ using equation (13).

**Step 2** Using the initial estimates $\pi_{il^*}$ maximize (17) to obtain $\beta$ (and $\lambda_0$, $C_i$) and $\hat{h}_l$

**Step 3** Update the estimate of $\pi_{il^*}$ using (10) and $\pi_l$ using equation (13).

**Step 4** Iterate between steps 2 and 3 until convergence.

3 Results

Example 1. In the EOP data there were approximately 4,000 herds at risk at the beginning of the study in 1989 and 2,100 at the end of the study in 2004 (due to amalgamation of herds). In all there were 4,891 distinct herds in the study. There were 2,293 events i.e. trade restrictions due to bovine TB in the study period. 991 herds had exactly one event while 502 had more than one. The Anderson-Gill form of the Cox model was fitted to the data. The final model had three significant covariates: log herd size as a time dependent covariate, a time varying previous history factor coded as the number of restrictions experienced by a herd up to that time, and a time varying treatment effect viz treatment*log(time). The same model was fitted again, now omitting the previous history covariate and using gamma frailty to model dependence between observations within a herd. A plot of the estimated frailty distribution can be seen in Figure 1. Herds with more than one event had larger estimated frailties which confirmed the model where event number / previous history was fitted as a fixed effect and found to be significant. In nonparametric terms two mass points can be identified from Figure 1. Therefore the nonparametric frailty was fitted using two mass points as outlined above. Using algorithm 1 it converged to a single mass point i.e. degenerate frailty. Algorithm 2 was extremely slow, moving in very small steps and was stopped, subject to further investigation.

Example 2. The rat data. The data are presented in Mantel et al.(1977) and used to illustrate the frailty computation in Therneau and Grambsch (2000) where further details can be found. There are 50 litters of 3 rats each, one of which received a potentially tumorigenic treatment. Survival time is time to development of a tumor. Therneau and Grambsch (2000) fitted a gamma frailty and found no evidence of a frailty effect. Fitting a nonparametric frailty with two mass points gave essentially the same result. The solution converged to one mass point i.e. degenerate frailty for both algorithms 1 and 2 above.
Example 3. Survival of kidney catheters. The data is presented in McGilchrist and Aisbett (1991) and also analyzed by Therneau and Grambsch (2000). The survival time is time to infection, at the point of insertion of the catheter, for kidney patients using portable dialysis equipment. There are 38 patients, each with exactly two observations. Therneau and Grambsch fitting a gamma frailty, identified one outlier in the data, subject number 21. With this subject removed the frailty was not important using a likelihood ratio test. Fitting nonparametric frailty with two mass points, algorithm 1 converged to mass points 1.06, 0.45 with associated probabilities 0.905, 0.095. Algorithm 2 converged to mass points 1.04, 0.05 with associated probabilities 0.962, 0.038.

Simulated data. Data were simulated from a Weibull model with a two mass point nonparametric frailty given by $h_1 = 1.6$ with probability $\pi_1 = 0.4$ and $h_2 = 0.6$ with probability $1 - \pi_1 = 0.6$. One hundred families of size 10 were generated from this distribution. Both algorithm 1 and 2 converged to similar solutions close to the true distribution after approximately 15 iterations. When 100 families of size 3 were generated from this distribution algorithm 1 converged to a degenerate solution and algorithm 2 had mass points (1.55, 0.66) with probabilities (0.38, 0.62), but it had not converged after 50 iterations. When 100 families were generated from this distribution, 70% of size 3 and 30% of size 1, algorithm 1 seemed to be converging to a degenerate solution after 100 iterations, and algorithm 2 had mass...
points (1.97,0.81) with probabilities (0.18,0.82), but it had not converged after 100 iterations - in particular $\pi_1$ was getting smaller.

4 Discussion

As is evident from the results, there are problems with the convergence of both algorithms. This is work in progress. A fast algorithm which is also reliable has yet to be determined. Further simulation is necessary to investigate both algorithms presented here. In addition we have only considered two mass points. Hartzel et al. (2001), Condon et al. (2004) and Aitkin et al. (2005) discuss the choice of number of mass points for nonparametric random effects distributions. They suggest the number of points be set initially to be low and then incremented. The optimal number is defined as that which maximizes the likelihood and does not result in duplicated mass points. This also needs to be considered.

A nonparametric frailty is appealing as it is based purely on the evidence of the data. In addition it may provide further information on the epidemiology of bovine TB and support for herd categorization.

Acknowledgments: Special thanks to the Centre for Veterinary Epidemiology and Risk Analysis (CVERA), University College Dublin, Ireland who provided the data for this study

References


Modelling mortality data on the Lexis diagram

James Kirkby\(^1\) and Iain Currie\(^1\)

\(^1\) Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, EH14 4AS, Scotland, jgk1@hw.ac.uk

Abstract: Currie, Durban & Eilers (2004) used 2-dimensional \(P\)-splines to smooth mortality data classified by age at death and year of death. In this paper we apply this model to data classified by age at death, year of death and year of birth. Discrete cohort effects are added to the model using a method similar to the overdispersion model of Perperoglou and Eilers (2006). This model allows us to decompose the mortality surface into (i) a smooth 2-dimensional surface in age and time and (ii) discrete cohort effects. We illustrate our remarks with the analysis of some German mortality data.

Keywords: Lexis diagram; mortality; overdispersion; \(P\)-splines.

1 Description of the data

We suppose we have census data on 1st January for years \(t = 1, \ldots, n_y\) which give the number of lives with ages between \(x\) and \(x + 1\) for ages labelled \(x = 1, \ldots, n_x\); these data are held in \(E = (e_{x,t})\), \(n_x \times n_y\). Let \(d_{x,t}\) be the number of deaths from the \(e_{x,t}\) lives in the calendar year \(t\) to \(t + 1\). These deaths can occur between ages \(x\) and \(x + 1\) (triangles of type \(A\) in Fig. 1), or between ages \(x + 1\) and \(x + 2\) (triangles of type \(B\)); triangles \(A\) and \(B\) are known as Lexis triangles (Carstensen & Keiding, 2005). Let \(D_{A} = (d_{A,x,t})\), \(n_x \times n_y\), hold the number of deaths in triangles of type \(A\) and \(D_{B} = (d_{B,x,t})\), \(n_x \times n_y\), hold the number of deaths in triangles of type \(B\). We assume that the three matrices \(E\), \(D_{A}\) and \(D_{B}\) are available. We seek estimates, \(E_{A}\) and \(E_{B}\), of the total times lived in triangles \(A\) and \(B\) respectively. Under the assumption that the \(d_{A,x,t}\) and \(d_{B,x,t}\) deaths are distributed uniformly over \(A\) and \(B\) respectively it is then straightforward to show that the expected times lived are

\[
E_{A} = \frac{1}{2} E - \frac{1}{6} D_{A} - \frac{1}{6} D_{B}, \quad E_{B} = \frac{1}{2} E - \frac{1}{2} D_{A} - \frac{1}{2} D_{B}.
\]  

For a life that dies in \(A\) the expected time of death is \(t + \frac{1}{3}\); similarly, for a life that dies in \(B\) the expected time of death is \(t + \frac{2}{3}\) and the expected age at death is \(x + \frac{4}{3}\). The situation is summarised in Fig. 1. We now model the number of deaths as follows

\[
d_{A,x,t} \sim P(e_{A,x,t} \mu_{x+2/3,t+1/3}) \quad (2)
\]

\[
d_{B,x,t} \sim P(e_{B,x,t} \mu_{x+4/3,t+2/3}) \quad (3)
\]
where $\mu_{x,t}$ is the hazard rate at age $x$ and time $t$. In this model the deaths and exposures in triangles of type A are located on a rectangular grid. The same remark applies to the data in triangles of type B so the whole data set consists of two interleaved grids. We seek simple forms for the function $\mu_{x,t}$ which decompose the mortality surface into (a) a smooth 2-dimensional surface and (b) discrete cohort and period effects. The smooth 2-dimensional surface represents underlying mortality while the discrete effects represent shocks to the surface caused, for example, by a cold winter or a flu epidemic (period effects) or by some effect at birth which influences the future mortality of that cohort. It is of particular interest to identify any such cohort effects.
2 A 2-dimensional smooth surface

We use the method of $P$-splines (Eilers & Marx, 1996). We take two rich bases of equally spaced cubic $B$-splines, one to cover the $x$-values from $1 + \frac{2}{3}$ to $n_a + \frac{4}{3}$ and another to cover the $t$-values from $1 + \frac{1}{3}$ to $n_y + \frac{2}{3}$. These bases give rise to marginal regression matrices $B_{A_a}$ and $B_{A_y}$ for the ages and years in triangles of type $A$. Since these data points lie on a grid, $B_A = B_{A_a} \otimes B_{A_y}$ is the regression matrix for a 2-dimensional surface; together with (2) we have defined a generalized linear model for data in triangles of type $A$. Similarly, $B_B = B_{B_a} \otimes B_{B_y}$ is a regression matrix for the data in triangles of type $B$. Finally $B = [B_A' : B_B']$ is the regression matrix of a surface for the complete data set. Smoothness of the fitted surface is ensured by marginal penalization and the model is fitted by penalized likelihood.

Fig. 2 shows a plot of the observed mortality surface (on the log scale) for some German mortality data (Statistisches Bundesamt). The obvious diagonal crests and troughs in Fig. 2 are indications of cohort effects. Fig. 2 also shows the surface fitted by the model described in the previous paragraph. The diagonal crests in the raw data are visible in the smooth surface, but
Fig. 2 suggests an alternative model where the individual cohorts effects are modelled as separate additive effects rather than as part of the underlying smooth mortality surface.

### 3 Adding cohort effects

We extend the basic smooth model to include cohort effects, as follows

\[
\eta = B a + C \gamma
\]  

(4)

where \( C \) is the design matrix for the individual cohort effects. A ridge penalty with smoothing parameter \( \kappa \) is applied to \( \gamma \). The ridge penalty maintains identifiability and the size of \( \kappa \) can be tuned to the observed cohort effects. The ridge penalty ensures that the smooth features of the data are described by the \( B \)-spline surface and only the additional variation caused by the cohort effects is absorbed by \( \gamma \).

For given values of the smoothing parameters, estimates of \( a \) and \( \gamma \) are obtained by solving:

\[
\begin{bmatrix}
B' W B + P_1 & B' W C \\
C' W B & C' W C + P_2
\end{bmatrix}
\begin{bmatrix}
a \\
\gamma
\end{bmatrix} =
\begin{bmatrix}
B' W \tilde{z} \\
C' W \tilde{z}
\end{bmatrix}
\]

(5)

where \( \tilde{z} \) is the usual working vector, and \( P_1 \) and \( P_2 \) are the appropriate difference and ridge penalty matrices. The values of the smoothing parameters are chosen by BIC.

Triangles \( A \) and \( B \) in the same cohort in Fig. 1 are modelled by a single parameter in model (4). However, lives in triangle \( A \) are born predominantly in the first half of the year while those in triangle \( B \) are born predominantly in the second half. Thus we can split the cohort effect into two parts.

**FIGURE 3. Smooth Surface: The smoothed surface obtained from the model with the linear predictor given in (6) plotted at data points of type A.**
extend model (4) and use separate parameters, one for each of the two types of triangle within a year of birth; the linear predictor becomes

$$\eta = Ba + C_A \gamma_A + C_B \gamma_B.$$  

(6)

This model gives a much improved fit (as measured by the BIC value). Figs. 3 and 4 show that this model successfully decomposes the mortality surface into a smooth 2-dimensional surface and a set of cohort effects. Period effects can be added in the same way as (4) but were not found to be significant for this data set.

There are some connections between our cohort models (4) and (6) and the overdispersion models of Perperoglou & Eilers (2006). These authors account for overdispersion by fitting an extra parameter for each data point. Their approach differs from the usual method of over-dispersion modelling which assumes the specification of the variance is not correct. In the standard approach estimation usually proceeds by using quasi-likelihood to re-specify the mean-variance relationship. The two approaches correspond to different underlying structures in the data: the Perperoglou & Eilers approach would be appropriate in the case of an underlying smooth trend shocked by individual random effects, while the second corresponds to a genuinely over-dispersed distribution for each data point. In one dimensional problems, it is difficult to distinguish between the two structures, and the quasi-likelihood approach has the advantage of being able to deal with under-dispersion. In higher dimensional problems, as here, we can distinguish between the two structures provided that the shocks follow some
kind of systematic pattern. In our case, the cohort patterns were clearly visible in the data, and the basic smooth model was not satisfactory, so an addition to the linear predictor was natural. In other situations close inspection of the data may lead to identification of additive effects that may otherwise be put down to over-dispersion.

It is also possible to fit model (4) by expressing it as mixed model. Re-parameterizing the $P$-spline model using the method given by Currie et al. (2006), the similarity between the system in (5) and the Fisher scoring algorithm given by Breslow & Clayton (1993) becomes clear. The smoothing parameters could then be optimized using ML or REML.

4 Conclusions

The richness of data on the Lexis Diagram allowed us to examine cohort effects in detail. For the data in our example we found strong evidence of cohort effects of two kinds: cohort effects, $\gamma_A$, associated with triangles of type A (early year births) and distinct cohort effects, $\gamma_B$, associated with triangles of type B (late year births).

The data set used in this paper was relatively small (12 years and 80 ages), and we were able to fit the model using standard matrix computations. If, for a bigger data set, the computations became unmanageable, an efficient algorithm to solve systems similar to (5) can be found in Perperoglou & Eilers (2006). Further gains are possible by taking advantage of the Kronecker product structure of the $B$-spline basis (Currie et al., 2006).

References


Statistisches Bundesamt, Various Tables (Data obtained through the Human Mortality Database, www.mortality.org, December 2005.)
A nonparametric multi-state model for the analysis of human sleep

Thomas Kneib\textsuperscript{1,2} and Andrea Hennerfeind\textsuperscript{1}

\textsuperscript{1} Department of Statistics, Ludwig-Maximilians-University Munich, Ludwigstraße 33, D-80539 München, Germany.
\textsuperscript{2} thomas.kneib@stat.uni-muenchen.de

Abstract: In this paper, we propose a class of nonparametric continuous time multi-state models for the analysis of human sleep data. The basic quantities in this model are multiplicative transition intensities that are specified in the spirit of the Cox model. Baseline intensities and time-varying effects are modelled using penalised splines. In addition, frailty terms are included to account for unobserved heterogeneity between the examined patients. Inference can be conducted either fully Bayesian using Markov chain Monte Carlo simulation techniques or empirically Bayesian based on a mixed model representation.

Keywords: Multi-state models; penalised splines; time-varying effects.

1 Multi-state models for human sleep data

The time-continuous process of human sleep is associated with a discrete state space of sleep stages. For example, in our application we will distinguish between the three states awake, rapid eye movement (REM), and non-rapid eye movement (Non-REM). In this study, the sleep stages of 39 patients were recorded every 30 seconds, providing sufficient information for a continuous time model. The basic quantities describing the dynamics of human sleep are the transition intensities between the sleep stages leading naturally to the consideration of multi-state models. While both Markov and Semi-Markov processes have been employed for the analysis of human sleep, they are not fully appropriate since they impose a very strict structure on the transition intensities. For example, the duration times between the transitions of a Markov process are independent and exponentially distributed with time-constant parameters. However, due to the changing dynamics of human sleep over night, more complex models embedded in the counting process framework seem to be recommended.

Let us introduce such multi-state models by means of a representative question in sleep research, i.e. whether a high level of cortisol is associated with an increased propensity for REM sleep. Figure 1 displays two typical realisations of individual sleep processes and the corresponding nocturnal cortisol secretion (measured on a coarser time level). Some association between the occurrence of REM phases and high cortisol level may be guessed from this picture but it is obviously quite difficult to detect associations and characteristic patterns by visual inspection without some kind of
smoothing and the simultaneous analysis of several sleep processes. In this case, individual sleeping habits and covariate effects such as gender differences have to be considered additionally. A multi-state model allows for the explicit description of the transition intensities in terms of time since sleep onset, individual sleeping behaviour and possible covariate effects such as hormonal secretion.

In our application, dynamics for transitions between REM and non-REM phases are of particular interest. Therefore, we chose a somewhat reduced model, schematically represented in Figure 2, which aggregates the tran-
FIGURE 2. Schematic representation of sleep stages and the transitions of interest.

sitions from awake to REM and Non-REM into one single transition and likewise for the reverse transition.

2 Specification of transition intensities

Inspired by the Cox model, we chose a multiplicative specification for the transition intensities of the process depicted in Figure 2:

\[
\begin{align*}
\lambda_{\text{AS},i}(t) &= \exp \left[ \gamma_0^{(\text{AS})}(t) + s_i \beta^{(\text{AS})} + b_i^{(\text{AS})} \right] \\
\lambda_{\text{SA},i}(t) &= \exp \left[ \gamma_0^{(\text{SA})}(t) + s_i \beta^{(\text{SA})} + b_i^{(\text{SA})} \right] \\
\lambda_{\text{NR},i}(t) &= \exp \left[ \gamma_0^{(\text{NR})}(t) + c_i(t) \gamma_1^{(\text{NR})}(t) + s_i \beta^{(\text{NR})} + b_i^{(\text{NR})} \right] \\
\lambda_{\text{RN},i}(t) &= \exp \left[ \gamma_0^{(\text{RN})}(t) + c_i(t) \gamma_1^{(\text{RN})}(t) + s_i \beta^{(\text{RN})} + b_i^{(\text{RN})} \right]
\end{align*}
\]

Due to the exponential link, the additive specification of the predictor on the right-hand side transforms to multiplicative effects on the transition intensities. Each of the \(\gamma_j(t)\) can be interpreted as a population averaged effect similar to the (log-) baseline hazard rate in the Cox model. The claimed effect of cortisol on the transitions between REM and Non-REM is allowed to vary smoothly over time by including varying coefficients terms \(c_i(t)\gamma_1^{(j)}(t)\) for the corresponding transitions. Since we are only interested in influences of high cortisol levels, \(c_i(t)\) is a dichotomised version of the original cortisol level at time \(t\), taking the value one for cortisol concentrations higher than 60 n mol/l and zero otherwise.

To account for individual-specific sleeping behaviour and gender differences, random effects \(b_i^{(j)}\) as well as a fixed gender effect \(s_i \beta^{(j)}\) are additionally
included. We also experimented with global random effects \( b_i \) independent of the transitions to allow for individuals with a generally increased or decreased tendency for changes between states. This leads to a multi-level random effects structure, where deviations from the global level are possible both on an individual and a transition-specific level. However, in our example it turns out that no global random effects are needed.

The baseline effects \( \gamma_0(t) \) and the time-varying effects \( \gamma_1(t) \) are modelled by Bayesian cubic P-splines with 20 inner knots (compare Brezger and Lang, 2006). The individual-specific effects are assumed to follow i.i.d. Gaussian distributions with transition-specific variances. Both assumptions lead to multivariate Gaussian priors of the form

\[
p(\xi_j | \tau_j^2) \propto \exp \left( -\frac{1}{2\tau_j^2} \xi_j' K_j \xi_j \right),
\]

where \( \xi_j \) represents either the B-spline coefficients describing a baseline or time-varying effect, or the collection of random effects for a specific transition. For penalised splines, the precision matrix \( K_j \) is formed by the cross-product of a difference matrix and the inverse variance \( 1/\tau_j^2 \) plays the role of the smoothing parameter in a frequentist setting. For random effects, the precision matrix is simply a unit matrix since the random effects are assumed to be independent and identically distributed. In analogy to the interpretation for penalised splines, the inverse variance of a random effect can also be interpreted as the penalty parameter of a ridge-type shrinkage penalty for the corresponding individual-specific effects.

### 3 Likelihood and inference

Embedding the present model into the counting process framework provides an explicit formula for the likelihood, which extends the likelihood of the Cox model to the multi-state situation. For each individual \( i, i = 1, \ldots, n \) the corresponding multi-state process can be represented by a set of counting processes \( N_{hi}(t) \), where \( h = 1, \ldots, k \) indexes the possible transitions in the model under consideration. Consequently, the jumps of the observed counting processes are defined by the transition times of the observed multi-state process. According to classical counting process theory, the intensity processes \( \alpha_{hi}(t) \) of the counting processes \( N_{hi}(t) \) are given by

\[
\alpha_{hi}(t) = Y_{hi}(t) \lambda_{hi}(t),
\]

where the transition intensities \( \lambda_{hi}(t) \) may be specified depending on covariates as in Section 2, and \( Y_{hi}(t) \) is an at-risk indicator, i.e. \( Y_{hi}(t) \) takes the value one if individual \( i \) is at risk for a type \( h \) transition at time \( t \) and zero otherwise.
Under mild regularity conditions this leads to the individual log-likelihood contributions

$$l_i = \sum_{h=1}^{k} \left[ \int_0^{T_i} \log(\lambda_{hi}(t)) dN_{hi}(t) - \int_0^{T_i} \lambda_{hi}(t) Y_{hi}(t) dt \right], \quad (2)$$

where $T_i$ denotes the time until which individual $i$ has been observed. The likelihood contributions can be interpreted similarly as in hazard rate models for survival times. The first term corresponds to contributions at the transition times which are given by the log-intensity for the observed transition evaluated at the particular time point. In survival models this term simply equals the log-hazard for uncensored observations evaluated at the survival time. The second term reflects cumulative intensities integrated over the waiting period between two successive transitions. The integral is evaluated for all transitions the corresponding person is at risk at during the current period. In survival models there is only one such transition (the transition from 'alive' to 'dead') and the integral is evaluated from the time of entrance to the study until the survival or censoring time.

These considerations yield an alternative representation of the likelihood, where each of the individual contributions is expressed in terms of transition indicators $\delta_{hi}(t)$ and observed transition times $t_{ij}$, $j = 0, \ldots, n_i$, where $t_{i0} = 0$ and $t_{in_i} = T_i$. The indicators $\delta_{hi}(t)$ take the value one if a transition of type $h$ is observed at time $t$ and zero otherwise, while the $t_{ij}$ are defined by the times at which the corresponding individual experiences a transition. This leads to the alternative log-likelihood formula

$$l_i = \sum_{j=1}^{n_i} \sum_{h=1}^{k} \left[ \delta_{hi}(t_{ij}) \log(\lambda_{hi}(t_{ij})) - Y_{hi}(t_{ij}) \int_{t_{i,j-1}}^{t_{ij}} \lambda_{hi}(t) dt \right], \quad (3)$$

which reveals more obviously the connection to the commonly known likelihood of hazard rate models in case of continuous survival times. Under the usual assumption of conditional independence the complete log-likelihood is given by the sum of the individual contributions. Note that the first integral in (2) reduces to a sum as shown in Equation (3) while the second integral has to be evaluated. When using splines of degree zero or one, explicit formulae for the integral can be derived. In general, however, a numerical integration technique has to be applied. We utilised the trapezoidal rule due to its simplicity but of course more sophisticated methods could also be used when desired.

As already noted before, formula (3) shows a close connection to the log-likelihood of a single-state Cox model for survival times. Accordingly, Bayesian inferential schemes developed for the Cox model can be adapted to the multi-state situation. Specifically, we considered the fully Bayesian approach based on Markov chain Monte Carlo simulation techniques proposed in Hennerfeind, Brezger and Fahrmeir (2006) and the empirical Bayes
FIGURE 3. Mixed model based estimates of the baseline transition intensities $\gamma_0(j)(t)$ (solid line) together with pointwise 95% credible intervals (dashed lines).

FIGURE 4. Mixed model based estimates of the time-varying effects $\gamma_1(j)(t)$ (solid line) together with pointwise 95% credible intervals (dashed lines).

approach based on a mixed model representation of nonparametric regression terms described in Kneib and Fahrmeir (2006). In the former, additional inverse Gamma hyperpriors are assigned to the variance parameters in (1) and an MCMC algorithm consisting of several Gibbs-
Metropolis-Hastings steps is constructed. In contrast, the variance parameters are treated as fixed unknown constants in the latter approach. Based on the mixed model formulation of penalised splines, the marginal likelihood of these variances can be derived yielding restricted maximum likelihood estimates.

Both approaches allow for joint estimation of the covariate effects and the corresponding smoothing parameters. Therefore, they enable a fully automated fit without the subjective choice of appropriate degrees of freedom for the nonparametric terms.

4 Results

Figure 3 displays some exemplary estimation results for the baseline transition intensities \( \gamma^{(j)}_0(t) \) obtained with the empirical Bayes approach. Obviously, all transition intensities show considerable variation throughout the night with lowest variation for the transition from REM to non-REM. An interesting observation is the cyclic behaviour of the intensity \( \exp[\gamma^{(NR)}_0(t)] \) which indicates an increased propensity for REM sleep every two hours starting approximately one hour after sleep onset. A similar cyclic pattern can be observed for the transition from sleep to awake. In addition, there seems to be an increasing trend in this transition over night with a sharp increase after approximately seven hours, which is, of course, perfectly reasonable from theoretical considerations.

The time-varying effects \( \gamma^{(j)}_1(t) \) for a high level of cortisol are visualised in Figure 4. For the transition from REM to non-REM there is only a modest effect and the horizontal line is included in the pointwise credible intervals. In contrast, the time-varying effect \( \gamma^{(RN)}_1(t) \) shows some additional cyclic pattern which is at least pointwise significantly different from zero. Since the cortisol level is generally low in the beginning of the night, the pointwise credible intervals are very wide for the corresponding part of the estimated effect. This is particularly true for the transition from Non-REM to REM, where the estimated effect is very small but with extremely wide credible intervals during the first two hours after sleep onset.

When comparing empirical Bayes and fully Bayesian estimates, results for the time-varying effects and gender differences are generally close to each other. Additional differences are encountered for the random effects. The mixed model approach detects almost no individual-specific heterogeneity while the MCMC approach identifies considerably more individual-specific variation. This is in accordance with our general observation that fully Bayesian inference has a tendency towards higher variance estimates, implying somewhat rougher estimates.
5 Discussion

We have presented a computationally feasible nonparametric approach to the analysis of multi-state duration data motivated by an application to human sleep. Transition intensities were specified in a multiplicative manner in analogy to the Cox model and all parameters were estimated jointly using either an empirical Bayes or a fully Bayesian approach. In the future, application to more complicated data structures including for example non-parametric covariate effects would be of interest. Of course, this requires a larger data base to make these effects identifiable. A methodological extension is the consideration of coarsened observations similar to interval censored survival data. This phenomenon is frequently observed in practice, in particular in medical applications where patients can be examined only at a prespecified fixed set of time-points. In a fully Bayesian approach, the augmentation of the true transition times in a data imputation step seems to be a promising idea. For the application of mixed model methodology the exact likelihood of a coarsened multi state model has to be evaluated leading to additional numerical difficulties.

Acknowledgments: We thank Judith Westermann for preparation of the data set provided by the Max-Planck-Institute of Psychiatry in Munich. This research was supported by the German Science Foundation, Collaborative Research Center 386 "Statistical Analysis of Discrete Structures".

References


Bias reduction and shrinkage in multinomial logit models

Ioannis Kosmidis

Abstract: A bias reduction method proposed by Firth (1993) is applied and evaluated in the case of baseline category logit models (see Agresti (2002) §7.1, Fahrmeir & Tutz (2001), §3.2). The intuitive argument that any results for this family of models should be direct generalizations of the corresponding results in the case of binary logistic regression, is stressed and validated. Also, we study the resultant effect on the estimates and their estimated standard errors. It is illustrated how the bias reduced estimates can be obtained using Iterative Generalized Least Squares (IGLS), and a fitting algorithm which is a modification of the standard fitting procedure is proposed. Quite important is that, as in the binary case, the method provides a solution to data separation issues in estimation.

Keywords: Logistic regression; baseline category logit models; bias reduction, shrinkage effects, iterative generalized least squares.

1 Background: The baseline category logit model and removal of the first order bias term

Bias correction in logistic regression has been studied by many authors (see for example Anderson & Richardson (1979), Copas (1988), Cordeiro and McCullagh (1991), and references therein). Bull et al. (2002) applied a method originally developed in Firth (1993) for removing the first-order bias term of the maximum likelihood estimates in multinomial logistic regression. However, an analytical study of the effect of the method to this family of models, and especially the way that the results in Firth (1992a, 1992b) generalize, is not provided in the latter paper and is pursued here.

1.1 The baseline category logit model

Assume multinomial response $Y$ with $k$ categories labelled as $1, 2, \ldots, k$, and let $\pi_1, \pi_2, \ldots, \pi_k$ be the corresponding category probabilities. In multinomial logistic regression the log-odds for category $a$ against category $b$ of the response is represented as follows:

$$\log \frac{\pi_a}{\pi_b} = (\alpha_a - \alpha_b) + (\beta_a - \beta_b)^T x,$$ (1)
with $x$ a vector of $p - 1$ covariate values, $\alpha_s \in \mathbb{R}$ and $\beta_s \in \mathbb{R}^{p-1}$. If, for identifiability reasons, we set $\alpha_h = 0$ and $\beta_h = 0$, where $h$ is the label of a reference category, the baseline category model (Agresti (2002) §7.1) results:

$$\log \frac{\pi_a}{\pi_h} = \eta_a = \alpha_a + \beta_a^T x, \quad a = 1, 2, \ldots, h - 1, h + 1, \ldots, k.$$  \hspace{1cm} (2)

Obviously, inferences using this model are invariant to the choice of the reference category since the only thing affected is the parametrization used. Thus, without loss of generality on what follows, we set as reference the $k$-th category of $Y$.

### 1.2 Removal of the first order bias term using Jeffreys prior

Maximum likelihood estimates for the above model are found to be biased away from the origin, as in the case of binary logistic regression. Especially, in small sample sizes the bias is considerable resulting in misleading, most of the time, inferences. So, a reduction of the bias would be attractive. Again, the effect of reducing the bias would be, naturally, a shrinkage of the estimates on their scale towards the origin and hence a reduction on the mean squared error (for the binary case see, for example, Heinze & Schumper (2002)).

The baseline category model is a multivariate generalized linear model in its natural form (Fahrmeir & Tutz (2001)) and is a direct generalization of ordinary binary logistic regression. Hence, as Firth (1993) showed, penalization of the likelihood $L$ with Jeffreys invariant prior will result to removal of the first order asymptotic bias term from the maximum likelihood estimates. The penalized likelihood is

$$L^M(\gamma) = L(\gamma) |I(\gamma)|^{1/2},$$ \hspace{1cm} (3)

with $I(\gamma)$ the Fisher information (which is equal to the observed information in the case of full exponential families in natural parametrization, as here). The bias-reduced estimates will be the roots of the modified score equations corresponding to (3).

### 2 Derivation of the modified score equations

Let $\gamma = (\alpha_1, \beta_1, \ldots, \alpha_q, \beta_q)^T$, with $q = k - 1$, be the vector of the $pq$ parameters of the model. Assume that we have observed $n$ pairs $(f_r, x_r)$ with $f_r = (f_{r1}, \ldots, f_{rk})$ the vector of observed frequencies for the corresponding levels of $Y_r$, and $x_r$ a $(p-1)$-dimensional covariate setting. The multinomial log-likelihood can be written as

$$l(\gamma; X, F) = \sum_r \sum_{s=1}^q f_{rs} \log \frac{\pi_{rs}}{\pi_{rk}} + \sum_r m_r \log(\pi_{rh}),$$ \hspace{1cm} (4)
with the natural parameters \( \log \pi_{rs} - \log \pi_{rk} \) modeled according (2) and \( m_r = \sum_s f_{rs} \). In what follows the matrix \( X \) with rows \( x_r \) is assumed to be of full rank. Writing \( Z_r = I_q \otimes (1, x_r^T) \) for the \( q \times pq \) model matrix, we can express (2) as

\[
\log \frac{\pi_{rs}}{\pi_{rk}} = \eta_{rs} = \sum_t \gamma_t z_{rst}, \quad s = 1, 2, \ldots, q, \tag{5}
\]

where \( z_{rst} \) should be thought as the \((s, t)\)-th element of \( Z_r \).

With the above notation the score vector is

\[
U(\gamma) = \sum_r U_r(\gamma) = \sum_r Z_r^T (f_r - m_r \pi_r). \tag{6}
\]

Also, for the current case, the Fisher information for \( \gamma \) takes the form

\[
I(\gamma) = Z^T \Sigma Z = \sum_r Z_r \Sigma_r Z_r^T \tag{7}
\]

with \( \Sigma \) being an \( nq \times nq \) block diagonal matrix with non-zero blocks the \( q \times q \) incomplete covariance matrices \( \Sigma_r = \text{var} f_r^{(-k)} \) and \( Z^T = (Z_1^T, \ldots, Z_n^T) \).

The vector \( f_r^{(-k)} \) corresponds to the first \( q \) components of \( f_r \).

From (3), the penalized log-likelihood has the form,

\[
l^*(\gamma) = l(\gamma) + \frac{1}{2} |I(\gamma)|. \tag{8}
\]

Differentiating with respect to \( \gamma_t \) we get the components of the modified score vector, giving rise to the score equations,

\[
0 = U^M_t(\gamma) = U_t(\gamma) + \frac{1}{2} \frac{\partial}{\partial \gamma_t} |Z^T \Sigma Z| \tag{9}
\]

\[
= U_t(\gamma) + \frac{1}{2} \text{trace} \left\{ (Z^T \Sigma Z)^{-1} (Z^T V_t Z) \right\} \tag{10}
\]

\[
= U_t(\gamma) + \frac{1}{2} \text{trace} \left\{ \Sigma^{-1} H V_t \right\}. \tag{11}
\]

In the above expressions, the block diagonal structure of \( \Sigma \) is inherited in its inverse. Also, \( H = \Sigma Z (Z^T \Sigma Z)^{-1} Z^T \) is an asymmetric form of the 'hat' matrix as is defined in the framework of multivariate generalized linear model (see Fahrmeir & Tutz §4.2.2 for definition and some details). Matrix \( H \) is a blocked \( nq \times nq \) matrix consisting of \( n^2 \) blocks (for clarity of notation we will denote by \( H_r \) the \( r \)-th diagonal block of \( H \)). Finally,

\[
V_t = \frac{\partial}{\partial \gamma_t} \Sigma
\]

is a block diagonal matrix with diagonal blocks \( V_{r|t} = \sum_{s=1}^q K_{r|s} z_{rst} \), where \( K_{r|s} \) is a \( q \times q \) symmetric matrix with \((u,v)\)-th element \( k_{r|s}^{uv} = \ldots\)
cum\((f_{rs}, f_{ru}, f_{rv})\), the incomplete third order cumulants of \(f_r\) (see, for example, McCullagh & Nelder (1989) p. 167 for analytic form of higher order cumulants of the multinomial distribution).

The modified score equations can thus be written as

\[
0 = U_i^M(\gamma) = U_i(\gamma) + \frac{1}{2} \sum_r \sum_{s=1}^q \text{trace} \left\{ \Sigma_r^{-1} K_{r|s} H_r^T \right\} z_{rst} .
\]  

(12)

After some algebra and replacing (6) in (12), the bias-reduced estimates are obtained by solving the equations,

\[
0 = U_i^M(\gamma) =
\sum_r \sum_{s=1}^q \left[ f_{rs} + \frac{1}{2} h_{rss} - \left( m_r + \frac{1}{2} \text{trace} \{H_r\} \right) \pi_{rs} - \frac{1}{2} \sum_{u=1}^q \pi_{ru} h_{rzu} \right] z_{rst} ,
\]  

(13)

where \(h_{rsu}\) is the \((s, u)\)-th element of \(H_r\).

When \(q = 1\), (2) reduces to the binary logistic regression model. Also, the matrices \(H_r\) correspond to the scalars \(h_r\), which are just the diagonal elements of the hat matrix in the univariate case. Thus, equation (13) reduces to

\[
0 = U_i^M(\gamma) = \sum_r \left[ w_r + \frac{1}{2} h_r - (m_r + h_r) \pi_r \right] z_{rt} ,
\]  

(14)

retrieving correctly the modified score equations for binary logistic regression (Firth 1992a, 1992b).

3 Iterative adjustments of the response

Considering \(H_r\) as matrices of constants with respect to \(\gamma\), the form of (13) suggests a pseudo-model for the response,

\[
f_r^{(-k)} + \frac{1}{2} \text{diag} \{H_r\} - \frac{1}{2} \text{trace} \{H_r\} \pi_r^{(-k)} - \frac{1}{2} H_r \pi_r^{(-k)} \sim \text{Multinomial} \left( m_r, \pi_r^{(-k)} \right)
\]  

(15)

where \(\text{diag} \{H_r\} = (h_{r11}, \ldots, h_{rqq})^T\) and \(f_r^{(-k)}\) and \(\pi_r^{(-k)}\) are, as before, the corresponding vectors of frequencies and probabilities for the first \(q\) categories of the \(r\)-th multinomial trial. The above non-standard expression for the multinomial distribution is justified by the fact that a multinomial distribution for a random variable with \(k\) categories can be viewed as a \((k-1)\)-dimensional distribution after the constraints \(\sum_s \pi_{rs} = 1\) and \(\sum_s w_{rs} = m_r\) are imposed.

However, \(H_r\) generally depends on \(\gamma\), except for some special cases, for example when the model is saturated. The usefulness of (15) is that it suggests a modified IGLS procedure for obtaining the bias-reduced estimates.
3.1 Saturated Models and Haldane Correction

Consider a saturated model of the form (2). This is described by \( nq \) parameters, as many as the number of the category probabilities (or more 'naturally' the log-odds) we are trying to estimate. For the saturated model the hat matrix \( H \) is the identity. Hence for this case the modified score equations take the form

\[
0 = \sum_t \sum_{s=1}^q \left[ f_{rs} + \frac{1}{2} \left( m_r + \frac{k}{2} \right) \pi_{rs} \right] z_{rst} - \sum_t \gamma_t z_{rst}.
\]

This suggests adding a half to each frequency and proceed using usual maximum likelihood on the modified response, retrieving the Haldane correction (Haldane 1956) introduced originally for avoiding singularities in the estimation of log-odds in sparse arrays, and producing the well-known bias-reducing modification for the empirical logistic transform. Parameter estimates in this case are obtained by solving

\[
\eta_{rs}(\gamma) = \sum_t \gamma_t z_{rst} = \log \left( \frac{f_{rs} + \frac{1}{2}}{f_{rk} + \frac{1}{2}} \right)
\]

with respect to \( \gamma \).

3.2 A solution to practical difficulties with separated datasets

Data separation in logistic regression has been studied extensively by Albert & Anderson (1984), Lesaffre & Albert (1989). With separated datasets, the maximum likelihood estimate is found to involve infinite components, causing in practice fitting procedures to fail to converge. Heinze & Schemper (2002) illustrated with 2 empirical studies that Firth’s bias reduction method provides a solution to the problem of separation in binary logistic regression. The same argument applies in the multinomial case. Obviously, (15) suggests that the initial multinomial counts are being adjusted by iteratively adding quantities that depend on the hat matrix and the category probabilities, thereby eliminating any initial singularities on the estimated log-odds and thus ensuring finite values for the estimated value of the linear predictor \( \eta_{rs} \) and consequently the parameter estimates themselves.

4 Proposed IGLS procedure for obtaining the bias-reduced estimates

For general models, we propose a modification of the original IGLS for obtaining maximum likelihood estimates. Assume that the current estimates are \( \gamma^{(c)} \). The proposed procedure will require two separate calculations for updating to \( \gamma^{(c+1)} \):
1. (i) Calculate $H_r(\gamma^{(c)}) = \Sigma_r(\gamma^{(c)}) Z_r \left[Z^T \Sigma_r(\gamma^{(c)}) Z_r\right]^{-1} Z_r^T$, for every $r = 1, \ldots, n$.

2. (ii) Treating $H_r = H_r(\gamma^{(c)})$ as known constants,

$$\gamma^{(c+1)} = \gamma^{(c)} + \left[Z^T \Sigma \left(\gamma^{(c)}\right) Z\right]^{-1} U^M(\gamma^{(c)})$$

where $U^M(\gamma^{(c)})$ has components as in (13), and can be viewed having the same functional form as the original score vector $U(\gamma^{(c)})$, with the first $q$ components of $f_r$ replaced by the ones specified by the pseudo-model (15) at the current step.

This procedure, which is a direct generalization of the procedure suggested by Firth (1992a), will generally converge in linear rate in contrast to the standard IGLS which converges in quadratic rate. The reason for this, is that only the first summand $I(\gamma) = Z^T \Sigma(\gamma) Z$ of minus the Jacobian of the modified scores is used. Note, that given that the Fisher information $I(\gamma)$ is positive definite, the above iteration will always cause an increase on the penalized likelihood and the uniqueness of the solution is guaranteed by the log-concavity of the penalized likelihood (see for example Bull et al 2002 §3.2 for an argument on concavity).

4.1 Starting values and computational issues

Staring values $\gamma^{(0)}$ obtained by the solution of (17) and the balanced choice $H_r^{(0)} = \frac{2}{n} I_q$ have been shown to be adequate in practice, dealing with any singularities in the observed counts. However, as Lesaffre & Albert (1989) proved, when either complete or quasi-complete separation occurs, at least a diagonal element of the inverse of the Fisher information diverges when maximizing the log-likelihood. This might cause numerical problems for the above fitting procedure, especially when the adjustments made are small, causing it to diverge or to locate stationary points away from the maximum argument. Implementing step-halving or golden section search, forcing the likelihood to increase at each iteration in such cases, eliminates this problem.

4.2 Estimated standard errors

Firth (1993) notes that the maximum likelihood estimate and the bias-reduced version of it agree in their asymptotic variances up to first order, both being the inverse of the Fisher information $I(\gamma)$. Thus, estimated standard errors for the bias reduced estimates can be obtained as byproduct of the suggested procedure: we can use the square roots of the diagonal elements of $(Z^T \Sigma(\gamma) Z)^{-1}$ at the last iteration.
TABLE 1. Analysis of the alligator food choice data.

<table>
<thead>
<tr>
<th>Logit Intercept</th>
<th>Size ≤ 2.3</th>
<th>Hancock</th>
<th>Oklawaha</th>
<th>Trafford</th>
</tr>
</thead>
<tbody>
<tr>
<td>I vs F</td>
<td>-1.55</td>
<td>1.46 (0.40)</td>
<td>-1.66 (0.61)</td>
<td>0.94 (0.47)</td>
</tr>
<tr>
<td></td>
<td>-1.49</td>
<td>1.40 (0.40)</td>
<td>-1.56 (0.60)</td>
<td>0.90 (0.47)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.9682)</td>
<td>(0.9546)</td>
<td>(0.9600)</td>
</tr>
<tr>
<td>R vs F</td>
<td>-3.31</td>
<td>-0.35 (0.58)</td>
<td>1.21 (1.19)</td>
<td>2.46 (1.12)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.9873)</td>
<td>(0.9685)</td>
<td>(0.9669)</td>
</tr>
<tr>
<td>B vs F</td>
<td>-2.09</td>
<td>-0.63 (0.64)</td>
<td>0.70 (0.78)</td>
<td>-0.65 (1.20)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.9789)</td>
<td>(0.9856)</td>
<td>(0.9806)</td>
</tr>
<tr>
<td>O vs F</td>
<td>-1.90</td>
<td>0.33 (0.45)</td>
<td>0.83 (0.56)</td>
<td>0.01 (0.78)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.9677)</td>
<td>(0.9812)</td>
<td>(0.9652)</td>
</tr>
</tbody>
</table>

1, invertebrate; R, reptile; B, bird; O, other; F, fish.

The first row of each cell corresponds to the ML estimate and its estimated standard error (parenthesized).
The second row of each cell corresponds to the bias-reduced estimate and its estimated standard error (parenthesized).
The curly bracketed number stands for the estimated coverage probability of the bias-reduced estimate, using two-tailed 95% CI’s and quantiles from the normal distribution. It is calculated after simulating 10000 samples under the current fit.

5 Alligator food choice data

In order to illustrate the effect of the bias reduction method on the estimates we apply it to the alligator food choice data in Agresti (2002) §7.1.2. The sample size is moderate (219 alligators studied) and the data are sparse (80 cells). We compute the bias reduced estimates for the best-fit-model according to Agresti. Therefore, after grouping the data over gender we model food choice against the lake that the alligators where captured and their size. The results in Table 1 illustrate the shrinkage towards the origin of the values of the bias-reduced estimates. Furthermore, their estimated standard errors are smaller in magnitude. Finally, the coverage of nominally 95% intervals based on the bias-reduced estimates is lower than that of corresponding intervals based on the MLE, but is still slightly conservative.

6 Generality of Results

Note that all the formulae from (7) to (12) and the proposed fitting procedure are valid and appropriate, correspondingly, for every multivariate generalized linear model in natural parametrization. Furthermore, if the dimension of the response is dropped to \( q = 1 \), they reduce to the corresponding formulae for generalized linear models with natural link function in Firth (1992a). However, general results and properties are not the purpose of the current paper and are not pursued here.
Acknowledgments: The author is more than grateful to professor David Firth for his valuable comments and suggestions.

References


Computationally efficient spatially-adaptive penalized splines

Tatyana Krivobokova$^1$, Ciprian M. Crainiceanu$^2$ and Göran Kauermann$^1$

$^1$ University of Bielefeld, Department of Economics, Postfach 100131, D-33501 Bielefeld, Germany, tkrivobokova@wiwi.uni-bielefeld.de, gkauermann@wiwi.uni-bielefeld.de

$^2$ Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, 615 N. Wolfe St. E3636 Baltimore, MD 21205, USA, ccrainic@jhsph.edu

Abstract: This paper proposes a numerically simple routine for locally adaptive smoothing. The idea is to link smoothing to hierarchical modelling and employ simple Laplace approximation to provide an easy and numerically handy fit. The approach is applied to discrete survival data where bivariate smoothing is carried out.

Keywords: Adaptive penalties; hierarchical modelling; mixed model.

1 Introduction

The recent years have seen an increasing use of penalized spline (P-spline) smoothing, thanks to its simple and practical idea. Tracing back to O’Sullivan (1986) the method has been made formal by Eilers & Marx (1996). A smooth unknown regression function is modelled in a parametric way with a high dimensional basis, imposing penalty on the basis coefficients to avoid overfitting. The link to linear mixed models allows using any linear mixed models software for penalized smoothing. However, the standard setting with a single penalization parameter fails if the function to be estimated is locally of varying complexity. To achieve spatial adaptivity we follow the idea of Crainiceanu, Ruppert & Carroll (2005) and model the smoothing parameter again with P-splines. Unlike their approach, however, we avoid numerically complex Monte Carlo routines but employ simple Laplace approximation.
2 Smoothly adaptive penalized splines for P-spline regression

2.1 Hierarchical penalty model

We introduce our method with the simple scatterplot smoothing model

\[ y_i \sim N(m(x_i), \sigma^2), \quad i = 1, \ldots, n, \]

where \( m(x) \) is a smooth function in the univariate metrical quantity \( x \). We replace \( m(x) \) for fitting by the penalized truncated polynomials

\[
m(x) = \beta_0 + x \beta_1 + \ldots + x^q \beta_q + \sum_{s=1}^{K_b} (x - \tau_s^{(b)})_+^q b_s =: X_b \beta + Z_b b,
\]

where \( \tau_1^{(b)}, \ldots, \tau_{K_b}^{(b)} \) are knots covering the range of \( x \) and \( (x - \tau_s^{(b)})_+^q \) is the truncated \( q \)-th order polynomial defined through \( (x - \tau_s^{(b)})_+^q \) if \( x - \tau_s^{(b)} > 0 \) and zero otherwise. The dimension \( K_b \) of the basis is chosen in a lush and generous manner and knots \( \tau_s^{(b)} \) are placed over the range of \( x \). To avoid overfitting we impose a penalty on the spline coefficients as a priori distribution

\[
b_s \sim N(0, \sigma_b^2(\tau_s))
\]

and model the variance as

\[
\sigma_b^2(\tau_s) = \exp[\gamma_0 + \tau_s^{(b)} \gamma_1 + \ldots + \tau_s^{(b)} \gamma_p + \sum_{t=1}^{K_c} (\tau_s - \tau_t^{(c)})_+^p c_t] =: \exp[X_c \gamma + Z_c c].
\]

Here \( \tau_1^{(c)}, \ldots, \tau_{K_c}^{(c)} \) is a second layer of knots covering the range of \( \tau_1^{(b)}, \ldots, \tau_{K_b}^{(b)} \). Again imposing a penalty on coefficients \( c \) as a priori distribution leads to the following hierarchical mixed model

\[
Y|b, c = X_b \beta + Z_b b + \epsilon, \quad \epsilon \sim N(0, \sigma^2 \epsilon I_n)
\]

\[
b|c \sim N(0, \Sigma_b), \quad \Sigma_b = \text{diag}[\exp[X_c \gamma + Z_c c]]
\]

\[
c \sim N(0, \sigma^2 c I_{K_c}).
\]

The likelihood for this model takes the form

\[
L(\beta, \gamma, \sigma^2_\epsilon, \sigma^2_c) = (2\pi)^{-\frac{n+K_c}{2}} \sigma_\epsilon^{-n} \sigma_c^{-K_c} \int_{R^{K_c}} \exp[-g(c)] dc, \tag{1}
\]

with

\[
g(c) = \frac{1}{2} \log |V_c| + \frac{c^T c}{2\sigma^2_c} + \frac{(Y - X_b \beta)^T V^{-1}_c (Y - X_b \beta)}{2\sigma^2_\epsilon}
\]

and \( V_c = I_n + Z_b \Sigma_b Z_b^T / \sigma^2_\epsilon \). The integral in (1) is not available analytically, which motivates a solution based on MCMC techniques. We however go a
different route via Laplace’s method, approximating the log-likelihood, up to a constant, by

\[-2l(\beta, \gamma, \sigma_e^2, \sigma^2) \approx n \log \sigma_e^2 + K_e \log \sigma_c^2 + \log |V_e(\hat{c})| + \log |I_{cc}(\hat{c})| + \hat{c}^T \hat{c}/\sigma_c^2 + (Y - X_\beta)^T V_e^{-1}(\hat{c})(Y - X_\beta)/\sigma_c^2, \]  

where

\[I_{cc}(\hat{c}) = 1/2 Z_c^T \text{diag}(v_{df}) Z_c + I_{K_c} \sigma_c^2 \]

and \(\hat{c}\) is the solution to

\[\frac{\partial g(\hat{c})}{\partial \hat{c}} = -1/2 Z_c^T \left\{\Sigma_{\hat{b}}^{-1} \hat{b}^2 - w_{df}\right\} + \frac{c}{\sigma_c^2} = 0.\]

Here \(\hat{b}\) is the prediction of \(b\) defined through

\[Z_c^T b = \sigma_c^2 \Sigma_{\hat{b}}^{-1} \hat{b} \]

and \(w_{df}\), \(v_{df}\) are \(K_b\) dimensional vectors containing diagonal elements of matrices \(A = Z_b^T Z_b(\sigma_c^2 \Sigma_{\hat{b}}^{-1} + Z_b^T Z_b)^{-1}\) and \(AA\), respectively. Assuming that weights \(v_{df}\) vary slowly or not at all as a function of \(\gamma\) we can estimate \(\gamma\) and \(c\) simultaneously, resulting the following iterated weighted least squares (IWLS) for estimation of parameter \(\theta = (\gamma^T, c^T)^T\)

\[\hat{\theta} = (W_c^{-1} v_{df}^2 W_c + D_c/\sigma_c^2)^{-1} W_c^{-1} v_{df} u, \]  

with \(W_c = (X_c, Z_c), D_c = \text{diag}(0_{p \times p}, I_{K_c})\) and \(u = W_c \theta + v_{df}^{-1} (\Sigma_{\hat{b}}^{-1} \hat{b}^2 - w_{df})\) as a working vector. Fixing now parameters \(\hat{\theta}\) and \(\hat{\sigma}_c^2 = \hat{c}^T / K_c\) in (2) provides the standard linear mixed model likelihood. Estimation can now be carried out by iterating between (3) and calls of any standard linear mixed models software for the remaining parameters updates. It should be noted that the estimation consists of two simple steps and is therefore numerically very fast.

2.2 Simulation and comparison with other univariate smoothers

To examine the performance of the above approach and to compare our results with those reported in Ruppert & Carroll (2000) and Baladandayuthapani, Mallick & Carroll (2005) we performed the following simulation study. For \(n = 400\) \(x\) equally spaced on \([0, 1]\) and independent \(\epsilon_i \sim N(0, 0.2^2)\) we considered the regression function

\[m(x) = \sqrt{x(1-x)} \sin \left(\frac{2\pi (1 + 2(9-4j)/5)}{x + 2(9-4j)/5}\right), \]

with \(j = 6\). We ran 500 simulations with \(K_b = 80\) and \(K_c = 20\). An exemplary fit (bold) together with confidence intervals (dashed) is shown on
We now generalize the ideas of the previous section to spatial smoothing

\[ y_i \sim N(m(x_i), \sigma_i^2), \quad i = 1, \ldots, n, \]

with \( x_i \in \mathbb{R}^2 \) and \( m(.) \) as a smooth function of 2 covariates. Following Crainiceanu, Ruppert & Carroll (2005) we use radial basis functions (for details see Ruppert, Wand & Carroll, 2003) and choose \( K_b \) knots \( \tau_1^{(b)}, \ldots, \tau_{K_b}^{(b)} \in \mathbb{R}^2 \). This defines the model matrices \( X_b \) with \( i \)-th row \( [1, x_i^T]_{1 \leq i \leq n} \) while the basis equals \( Z_b = Z_{K_b} \Omega_{K_b}^{-1/2} \) where \( Z_{K_b} = [\|x_i - \tau_s^{(b)}\|^2 \log \|x_i - \tau_s^{(b)}\|_{1 \leq s \leq K_b, 1 \leq i \leq n}] \) and \( \Omega_{K_b} = [\|\tau_t^{(b)} - \tau_s^{(b)}\|^2 \log \|\tau_t^{(b)} - \tau_s^{(b)}\|_{1 \leq s,t \leq K_b}] \) with

\[ \hat{\theta} = (X_b^T X_b)^{-1} X_b^T y. \]

Figure 1. Estimated regression function \( m(x) \) (bold) with confidence intervals (dashed) and true function.

3 Extension to spatial smoothing

We now generalize the ideas of the previous section to spatial smoothing

\[ E((f(x) - \hat{f}(x))^2) \]

with the expectation being replaced by the mean of the simulations. For better visual impression we show a simple smoother (thick line) for the latter. The average MSE over all \( x \)'s (AMSE) equals 0.0033, which is comparable with 0.0027 reported in Baladandayuthapani, Mallick & Carroll (2005 and 0.0026 of Ruppert & Carroll (2000). We computed also the coverage probabilities of the 95% confidence intervals over all 500 simulated datasets. Right plot in Figure 2 shows smoothed pointwise coverage probabilities. For small values of \( x \leq 0.1 \), i.e. in the region with low signal-to-noise ratio, there is clear undercoverage but beyond 0.1, say the coverage probability exceeds 95% being slightly conservative. The average coverage probability results to 94.95%.
\[ ||.|| \] denoting the Euclidean norm in \( \mathbb{R}^2 \). Including penalties and using the link to linear mixed model we get
\[
Y | b = X_b \beta + Z_b \epsilon + \epsilon, \epsilon \sim N(0, \sigma^2 \epsilon I_n),
\]
\[
b \sim N(0, \Sigma_b). \quad (4)
\]

Local adaptive smoothing is now implemented by allowing coefficients \( b \) to have locally varying variance. Like above we set subknots \( \tau^{(b)}_{1}, ..., \tau^{(b)}_{K_b} \in \mathbb{R}^2, K_b < K \), and define matrices \( X_b \) and \( Z_b \) similarly to the corresponding definition of matrices \( X_b \) and \( Z_b \) that is \( X_{s}^{n} = [1, (\tau_{s}^{(b)})^{T}]_{1 \leq s \leq K_b} \), \( Z_{c} = Z_{K_c} \Omega_{K_c}^{-1/2} \) with \( Z_{K_c} = \| \tau_{s}^{(b)} - \tau_{t}^{(c)} \|^{2} \log \| \tau_{s}^{(b)} - \tau_{t}^{(c)} \|_{1 \leq s \leq K_b, 1 \leq t \leq K_c} \) and \( \Omega_{K_c} = \| \tau_{s}^{(c)} - \tau_{t}^{(c)} \|^{2} \log \| \tau_{s}^{(c)} - \tau_{t}^{(c)} \|_{1 \leq s \leq K_c} \), \( 1 \leq t \leq K_c \) where the \( x \) covariates are replaced by knots \( \tau^{(b)} \) and the knots are replaced with subknots \( \tau^{(c)} \). The model is completed by adding to (4) the hierarchical structure
\[
\Sigma_b = \text{diag}[\exp(X_c \gamma + Z_c \epsilon)], \epsilon \sim N(0, \sigma^2_c I_{K_c}).
\]

Estimation can now be carried out analogously to above.

4 R Package “AdaptFit”

To implement our approach we developed an R package. We took advantage of the R package “SemiPar”, written by M.P. Wand to accompany the book Rupert, Wand & Carroll (2003). To perform adaptive smoothing we integrate the Fisher scoring procedure for \( \theta \) with updates of the remaining parameters by subsequent calls of function “spm” of the package “SemiPar”.

FIGURE 2. Left plot: pointwise MSE for 500 simulated datasets with function \( m_1(x) \). Solid line shows a smoother of the points. Right plot: smoothed pointwise coverage probabilities of 95% confidence intervals for 500 simulated datasets with function \( m_1(x) \).
The current version of our package “AdaptFit” with the function “asp” is available at www.wiwi.uni-bielefeld.de/~statistics/software/software.html. In general the usage of “asp” is similar to that of function “spm”, examples are provided within the package.

5 Conclusion

We demonstrated how local adaptive smoothing can be easily carried out by formulating penalties on spline coefficient as hierarchical mixed model. The major contribution was to show how simple Laplace approximation of the marginal likelihood allows to fit such models relatively easy without MCMC methods. The approach is implemented in the R package "AdaptFit".

References


A model for combining evidence in several $t$-statistics to estimate an overall standardized effect

Elena Kulinskaya$^1$ and Robert G. Staudte$^2$

1 Statistical Advisory Service, Imperial College, Sir Alexander Fleming Building, South Kensington Campus, London, SW7 2AZ, UK, e.kulinskaya@imperial.ac.uk
2 Department of Statistical Science, La Trobe University, Melbourne, Australia 3086, r.staudte@latrobe.edu.au

Abstract: Given several related experiments measuring related but unequal fixed effects, it is desired to combine the results and estimate a representative overall effect. In the context of one-sample normal models, the Student $t$-statistic can be variance stabilized, which leads to a confidence interval for the standardized effect. The overall standardized effect for several such studies is defined as the one which transforms into a weighted average of the respective transformed standardized effects, and a confidence interval for it is obtained. The method is also illustrated for two-sample normal models based on the Welch $t$-statistic.

Keywords: Confidence intervals; fixed effects; meta-analysis; variance stabilizing transformations; Welch $t$-statistic.

1 Introduction

Given $K$ independent but related experiments with $n_k$ observations in the $k$th leading to an estimate $\hat{\theta}_k$ of an unknown effect $\theta_k$, it is often desired to combine these results to estimate a representative overall effect $\theta$. For example, in one-sample experiments modeled by the normal distribution with unknown mean $\mu_k$ and variance $\sigma_k^2$ in the $k$th study, there may be interest in the unknown effects $\theta_k = \mu_k - \mu_0$ or the standardized effect $\delta_k = (\mu_k - \mu_0)/\sigma_k$. Two models often proposed for an overall effect are the fixed effects model, which assumes all $\theta_k = \theta$ and the random effects model, which assumes the $\theta_k$’s are themselves a random sample from a normal population with mean $\theta$ and variance $\gamma^2$, say. In both these models the overall representative effect $\theta$ is introduced by making very strong assumptions, which are unlikely to hold in practice. It would be very surprising that the $\theta_k$’s were all equal or that they were a random sample from some distribution while at the same time the population variances $\sigma_k^2$ are fixed and even assumed equal.
In standard meta-analytic theory, an overall weighted effect is defined by
\[ \theta = \sum_k w_k \theta_k, \]
with weights \( w_k \propto n_k / \sigma^2_k \), but this approach suffers from
the fact that these weights are difficult to estimate, and thus estimators of \( \theta \)
are only accurate for very large sample sizes relative to those needed to
estimate the individual \( \theta_k \).

Alternatively, one can transform statistics such as \( S_k = (\bar{\mu}_k - \mu_0) / \hat{\sigma}_k \) to
\( T_k = \sqrt{n_k} m(S_k) \) which is approximately normally distributed with mean \( \tau_k \) and variance 1, abbreviated \( T_k \sim N(\tau_k, 1) \). A linear combination of these
variance stabilized \( T_k \)'s then leads readily to a confidence interval for the
same linear combination \( \tau \) of the \( \tau_k \)'s; and, this interval can be transformed
back into a confidence interval for a parameter which represents an overall effect. More precisely, suppose that each \( \tau_k = \sqrt{n_k} m(\delta_k) \) and define \( \tau = \sum_k \sqrt{w_k} \tau_k \), where \( w_k = n_k / n \) with \( n = \sum_k n_k \). Then \( T = \sum_k \sqrt{w_k} T_k \sim N(\tau, 1) \) so a 100(1 - \( \alpha \))% confidence interval for \( \tau \) is \( T \pm z_{1-\alpha/2} \). Now we can write \( \tau / \sqrt{n} = \sum_k w_k \tau_k / \sqrt{n_k} = \sum_k w_k m(\delta_k) \). Thus \( \tau / \sqrt{n} \) is a weighted
average of the \( m(\delta_k) \) and because \( m \) is monotone increasing in \( \delta \) there is
a unique \( \delta = m^{-1}(\tau / \sqrt{n}) \) which we call the overall standardized effect. It
is the \( \delta \) which transforms via \( m \) into the weighted average of the \( m(\delta_k) \). A
100(1 - \( \alpha \))% confidence interval for \( \delta \) is clearly given by

\[
\left[ m^{-1}\left( (T - z_{1-\alpha/2}) / \sqrt{n} \right), m^{-1}\left( (T + z_{1-\alpha/2}) / \sqrt{n} \right) \right].
\]

2 Variance stabilizing transformations

While variance stabilization is straightforward in principle (Johnson, Kotz
and Kemp, p.55, 1993), in practice it is usually much harder. In the one-
sample normal model, Azorin (1953) shows that the Student-\( t \) statistic
\( t_{nu} = \sqrt{n} (\bar{x}_n - \mu_0) / s_n \) with \( \nu = n - 1 \) degrees of freedom can be variance
stabilized by a transformation which is asymptotic to \( \sqrt{\tau} m(t_{nu} / \sqrt{\tau}) \), where
\( m(x) = \sqrt{2} \sinh^{-1}(x / \sqrt{2}) \). This transformed \( t_{nu} \) is biased for \( \tau = \sqrt{n} m(\delta) \),
where \( \delta = (\mu - \mu_0) / \sigma \) is the unknown standardized effect. A bias corrected
version called \( T \) satisfies the normal approximation \( T \sim N(\tau, 1) \) and leads
to very reliable 95% confidence intervals for \( \delta = m^{-1}(\tau / \sqrt{n}) \) for sample
sizes as small as 10 and all \( |\delta| < 2 \).

In the case of two independent normal samples of sizes \( m, n \), all parameters
unknown, Kulinskaya and Staudte (2006) have derived a variance stabilizing
transformation for the Welch (1948) statistic \( t_{Welder} \) with \( \nu \) degrees of freedom.
Letting \( N = m + n \), and \( \xi = N / \nu \), the resulting evidence satisfies
\( T \sim N(\tau, 1) \) approximately, with \( \tau = \sqrt{2N / \xi} \sinh^{-1}(\sqrt{\xi} \delta / \sqrt{2}) \). Simulations
demonstrate the reliability of confidence intervals for the standardized
effect \( \delta = (\mu_2 - \mu_1) / \sigma \), where \( \sigma^2 / N = \sigma_1^2 / m + \sigma_2^2 / n \), for sample sizes as
small as \( m = n = 5 \) when \( |\delta| < 1 \) and the ratio of variances is between 1/4 and 4.
TABLE 1. Two-sample data for each of seven studies comparing drop in systolic blood pressure for treated patients undergoing a weight-loss regime (summarized by \( n, \bar{y}, s_2 \)) with control patients not undergoing a weight-loss regime (summarized by \( m, \bar{x}, s_1 \)).

<table>
<thead>
<tr>
<th>Study</th>
<th>( n_k )</th>
<th>( \bar{y}_k )</th>
<th>( s_{2k} )</th>
<th>( m_k )</th>
<th>( \bar{x}_k )</th>
<th>( s_{1k} )</th>
<th>( \hat{\delta}_k )</th>
<th>( T_k )</th>
<th>( U_k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haynes</td>
<td>27</td>
<td>4.8</td>
<td>13.8</td>
<td>24</td>
<td>0.2</td>
<td>13.8</td>
<td>-0.35</td>
<td>-1.79</td>
<td>-0.75</td>
</tr>
<tr>
<td>MacMahon</td>
<td>20</td>
<td>13.3</td>
<td>8.1</td>
<td>18</td>
<td>7.4</td>
<td>8.1</td>
<td>1.64</td>
<td>6.27</td>
<td>+1.04</td>
</tr>
<tr>
<td>Croft</td>
<td>66</td>
<td>11.0</td>
<td>17.1</td>
<td>64</td>
<td>4.0</td>
<td>15.7</td>
<td>0.64</td>
<td>5.06</td>
<td>+0.39</td>
</tr>
<tr>
<td>Andersson</td>
<td>10</td>
<td>4.0</td>
<td>15.3</td>
<td>9</td>
<td>-3.0</td>
<td>13.5</td>
<td>0.26</td>
<td>0.82</td>
<td>-0.36</td>
</tr>
<tr>
<td>Jalkanen</td>
<td>24</td>
<td>8.0</td>
<td>20.4</td>
<td>25</td>
<td>15.0</td>
<td>16.5</td>
<td>0.39</td>
<td>1.90</td>
<td>-0.01</td>
</tr>
<tr>
<td>Kawamura</td>
<td>5</td>
<td>9.8</td>
<td>7.1</td>
<td>5</td>
<td>2.5</td>
<td>5.1</td>
<td>1.38</td>
<td>2.73</td>
<td>+0.35</td>
</tr>
<tr>
<td>Gordon</td>
<td>19</td>
<td>12.5</td>
<td>6.3</td>
<td>14</td>
<td>9.9</td>
<td>6.4</td>
<td>1.98</td>
<td>7.03</td>
<td>+1.30</td>
</tr>
</tbody>
</table>

3 Examples

Seven studies are selected from the review by Mulrow et al. (2004) in each of which the drop in systolic blood pressure following a weight reducing diet for a group of patients was compared to that of a control group; the data are displayed in Table 3.1. See the Cochran Review website www.nicsl.com.au for more background and related material.

3.1 One-sample analysis

For the treated patients the standardized effects for each study are shown in Column 8 of Table 3.1, the Student-\( t \) statistic in Column 9 and 95% confidence intervals for the respective \( \delta_k \) in Column 10. These intervals show that \( \delta_k = 0 \) would be rejected in favor of two-sided alternatives at level 0.05 for studies \( k = 2, 3, 6 \) and 7.

If we combine the \( T_k \)’s to obtain \( T = \sum_k \sqrt{w_k} T_k = 8.29 \) as described in the previous section we obtain a point estimate \( \hat{\delta} = m^{-1}(T/\sqrt{n}) = 0.65 \) of the overall effect \( \delta = a^{-1}(\tau/\sqrt{n}) \) and a 95% confidence interval for \( \delta \) of \([-0.49, +0.83]\). Thus even though only 4 of the seven studies indicate sufficient evidence against a null effect at level 0.05, and the contrary studies are included in the overall effect, the combined evidence rejects the null hypothesis that \( \delta = 0 \) at level 0.05.

3.2 Two-sample analysis

Next we consider comparing treatment with control results using a variance stabilizing transformation of the two-sample Welch statistic derived in Kulinskaya and Staudte (2006). Here \( N_k = m_k + n_k \), the effect in the \( k \)th study is \( \theta_k = \mu_{2k} - \mu_{1k} \) and the standardized effect is \( \delta_k = \theta_k/\sigma_k \), where \( \sigma_k^2/N_k = \sigma_{1k}^2/m_k + \sigma_{2k}^2/n_k \). The results are displayed in Table 3.2.
TABLE 2. Values of the estimated effects $\hat{\theta}_k$, its standard error $\hat{\sigma}_k/\sqrt{N_k}$, the $t_{Welch,k}$ statistic, the variance stabilized statistic $T_k = T(t_{Welch,k})$ and respective 95% confidence intervals $[L_k, U_k]$ for each $\delta_k$.

<table>
<thead>
<tr>
<th>k</th>
<th>N_k</th>
<th>$\hat{\theta}_k$</th>
<th>$\hat{\sigma}_k/\sqrt{N_k}$</th>
<th>$t_{Welch,k}$</th>
<th>$T_k$</th>
<th>$L_k$</th>
<th>U_k</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>51</td>
<td>-5.0</td>
<td>3.87</td>
<td>-1.29</td>
<td>-1.26</td>
<td>-0.46</td>
<td>+0.10</td>
</tr>
<tr>
<td>2</td>
<td>38</td>
<td>5.9</td>
<td>2.63</td>
<td>2.24</td>
<td>2.16</td>
<td>+0.03</td>
<td>+0.70</td>
</tr>
<tr>
<td>3</td>
<td>130</td>
<td>7.0</td>
<td>2.88</td>
<td>2.43</td>
<td>2.40</td>
<td>+0.04</td>
<td>+0.39</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>7.0</td>
<td>6.61</td>
<td>1.06</td>
<td>1.00</td>
<td>-0.22</td>
<td>+0.71</td>
</tr>
<tr>
<td>5</td>
<td>49</td>
<td>-7.0</td>
<td>5.31</td>
<td>-1.31</td>
<td>-1.29</td>
<td>-0.47</td>
<td>+0.10</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>7.3</td>
<td>3.91</td>
<td>1.87</td>
<td>1.61</td>
<td>-0.11</td>
<td>+1.37</td>
</tr>
<tr>
<td>7</td>
<td>33</td>
<td>2.6</td>
<td>2.24</td>
<td>1.16</td>
<td>1.18</td>
<td>-0.15</td>
<td>+0.55</td>
</tr>
</tbody>
</table>

Only the 2nd and 3rd studies are of level 0.05 significance in testing $\delta_k = 0$ against $\delta_k \neq 0$.

The total number of observations is $N = 330$, the total degrees of freedom is $307.9$, so $\xi = 1.072$. The variance stabilized Welch statistics are combined to obtain $T = 2.12$ and an estimate of the overall standardized effect is $\hat{\delta} = 0.18$, which lies between the two negative and five positive standardized effects obtained in the 7 studies. While it is small, the 95% confidence interval for the overall effect $\delta$ is given by $[L, U] = [0.01, 0.23]$, and hence there is level 0.05 significance against $\delta = 0$ in favor of $\delta \neq 0$. Thus as in the one-sample case conflicting results in the 7 studies can be combined, but in the end these results reject the null hypothesis at level 0.05 in the direction of a drop in systolic blood pressure for those undergoing a weight loss program over those who do not undergo it.

References


Bayesian multi-dimensional density estimation with P-splines

Philippe Lambert$^1$ and Paul H. C. Eilers$^2$

$^1$ Institut de statistique, Université catholique de Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium, lambert@stat.ucl.ac.be
$^2$ Department of Medical Statistics, Leiden University Medical Centre, The Netherlands

1 Introduction

Density estimation is a neglected topic in modern Bayesian statistics. Heavy computation with sophisticated algorithms is the standard, but results are mostly presented as kernel-smoothed (one-dimensional) distributions, using default settings, or just as histograms. In the Bayesian literature, density estimation has been treated as a step-child; it is not considered explicitly in standard reference books. Our goal here is to fill in the gaps, using a combination of polytomous logistic regression, penalized splines and efficient simulation.

We model the logarithm of the density as a sum of scaled B-splines. This is similar to work by Kooperberg and Stone (1991). However, we do not try to optimize the number and positions of the knots that define the B-splines. Instead, we start out with many equally spaced knots, defining a basis that is ”too rich”: it provides more flexibility than needed. To get the desired smoothness, a difference penalty is put on the coefficients and the weight of this penalty is tuned to the data. This is the P-spline approach, advocated by Eilers and Marx (1996) and inspired by the work of O’Sullivan (1988).

In a Bayesian setting the penalty is the logarithm of a prior density of the differences of the coefficients. Efficient simulation is possible with the Langevin-Hastings algorithm (Roberts and Rosenthal, 1998) and proper rotation of the parameter vector. This approach has shown its value in hazard estimation in survival models with varying coefficients (Lambert and Eilers, 2005). In some sense, one-dimensional density estimation is simplified hazard estimation (without covariates), so we could stop here.

However, we extend our approach to multidimensional density estimation, making use of recently developed fast methods for weighted regression on tensor product basis functions with data on grids.

The spline coefficients are found by polytomous logistic regression associated to counts in a histogram with narrow bins. Polytomous logistic regression can be extended to multidimensional histograms, using tensor
Bayesian multi-dimensional density estimation

products of B-splines as basis functions. The difference penalties can be extended to multiple dimensions too, so P-splines can be generalized to this setting. However, straightforwardly constructing the basis matrix and performing the weighted regressions, leads to problems in memory use and computation time. Eilers et al. (2006) present an algorithm in which the multidimensional basis matrix is avoided completely. The computations are rearranged in such a way that one works along each dimension separately. This saves orders of magnitude in memory use and computation time. The algorithm has been used successfully for smoothing (and extrapolation) of large mortality tables (Currie et al., 2004). The underlying array algorithms are presented more formally by Currie et al. (2006).

2 Univariate density smoothing

Assume that a sample \( \{y_j, j = 1, \ldots, n\} \) of a random variable \( Y \) has been observed and that estimation of the density \( f_Y \) of \( Y \) is of interest. Following Eilers and Marx (1996), we propose to tackle the problem by starting from the histogram associated to a large number \( I \) of bins with equally spaced limits. This requires the specification of a compact interval \([y_{\text{min}}, y_{\text{max}}]\) over which most of the probability mass is expected to be found. Let \( x_i \) denote the center of the \( i \)th bin and \( n_i \) be the number of observations in that bin of width \( \Delta \). Then, it is well known that

\[
(N_1, \ldots, N_I) \sim \text{Mult}(\pi_1, \ldots, \pi_I) \quad \text{where} \quad \pi_i = \int_{x_i - \Delta/2}^{x_i + \Delta/2} f(z) \, dz \approx f(x_i) \Delta.
\]

Consider a basis \( \{b_k(\cdot) : k = 1, \ldots, K\} \) of cubic B-splines associated to equidistant knots on \([y_{\text{min}}, y_{\text{max}}]\). If \((B)_{ik} = b_k(x_i)\) denotes the \( I \times K \) matrix associated with that basis evaluated at the bin midpoints, then a possible model for \( \pi = (\pi_1, \ldots, \pi_I)' \) is the polytomous logistic regression

\[
\pi_i = \frac{\exp{\eta_i}}{\sum_{k=1}^{K} \exp{\eta_k}} \quad \text{with} \quad \eta_i = \sum_k b_{ik} \phi_k \quad \forall \ i
\]

with an identifiability constraint such as \( \sum_k \phi_k = 0 \). A large number of knots (say 20) is recommended to give enough flexibility with the approximation; however, that flexibility should be counterbalanced by a roughness penalty to give a smooth estimate of the density.

In a Bayesian setting, a roughness penalty translates into a prior distribution on the spline coefficients (see e.g. Lambert and Eilers (2005) in survival analysis).

3 Bivariate density smoothing

Assume that a random sample \( \{(y_{1k}, y_{2k}) : k = 1, \ldots, n\} \) of a bivariate random variable \((Y_1, Y_2)\) has been observed and that estimation of the
bivariate density \( f_{Y_1,Y_2} \) is of interest. The same ideas as in the 1D case can be used successfully. Consider for simplicity that most of the probability mass is within a rectangle and that this region can be subdivided as a \( I \times J \) grid of rectangles of equal areas. If \((x_{1i}, x_{2j})\) and \(n_{ij}\) denote, respectively, the midpoint and the number of data in cell \((i,j)\), then

\[
(N_{11}, \ldots, N_{IJ}) \sim \text{Mult}(\pi_{11}, \ldots, \pi_{IJ})
\]

where

\[
\pi_{ij} = \int \int_{\text{cell } (i,j)} f_{Y_1,Y_2}(z_1, z_2) \, dz_1 \, dz_2 \approx f_{Y_1,Y_2}(x_{1i}, x_{2j}) \Delta_1 \Delta_2
\]

Consider two bases of cubic B-splines \( \{b_k(\cdot) : k = 1, \ldots, K\} \) and \( \{b_l(\cdot) : l = 1, \ldots, L\} \) associated with equidistant knots along both axes. If \((\bar{B})_{ik} = \bar{b}_k(x_{1i})\) and \((B)_{jl} = b_l(x_{2j})\) denote the \(I \times K\) and \(J \times L\) matrices associated to these bases at their respective bin midpoints, then a possible model for the \(I \times J\) matrix of probabilities \((\Pi)_{ij} = \pi_{ij}\) is the polytomous logistic regression

\[
\pi_{ij} = \frac{e^{\eta_{ij}}}{e^{\eta_{i1}} + \ldots + e^{\eta_{iJ}}}
\]

where \(\eta_{ij} = \sum_k \sum_l \bar{b}_{ik} b_{jl} \phi_{kl} = (\bar{B}\phi B')_{ij}\).

We propose to force smoothness by considering a prior distribution on the \(r\)th order differences of the successive B-splines coefficients associated to each row and to each column of \(\Phi\). One can show that such a prior is

\[
p(\Phi|\tau_r, \tau_c) \propto \sqrt{d(P)} \exp \left\{ -\frac{1}{2} \text{vec}(\Phi)^t P \text{vec}(\Phi) \right\}
\]

where \(P = \tau_r P_r \otimes I_K + \tau_c I_L \otimes P_c\) and \(d(P)\) denotes the product of non-zero eigenvalues of \(P\). It is a function of row and column penalty weights \(\tau_r\) and \(\tau_c\).

4 The Langevin-Hastings algorithm

4.1 The basic algorithm

Several algorithms based on MCMC can be set up to explore the posterior. Conditionally on the penalty parameters, we are left with the well-studied problem of exploring the posterior of regression parameters in generalized linear models (see e.g. Gamerman, 1997; Brezger and Lang, 2006). The roughness penalty parameters, given the spline parameters, have identified conditional posterior distributions. Hence, a Gibbs sampler is easy to set up.

However, given the potentially large number of B-spline parameters, we think that the Metropolis-adjusted Langevin algorithm (MALA, Roberts
Bayesian multi-dimensional density estimation

and Tweedie, 1996) is better suited as it just requires the computation of the log-posterior and of its gradient at each iteration: no potentially large precision matrix must be computed as in the Bayesian version of the Iterative Weighted Least Squares (IWLS) algorithm involved in the last two references. Moreover, if the marginal posterior of the B-spline parameters is considered, no sampling from the roughness penalty parameters is required (Lambert, 2005).

The MALA algorithm builds McMC chains with proposals relying on the gradient of the log posterior distribution at the current state. More precisely, if \( p(\theta | y) \) is the posterior distribution and \( \theta^t \in \mathbb{R}^K \) the state of the chain at iteration \( t \), then the proposal \( \theta \) for the next state is obtained by a random generation from the \( K \)-variate normal distribution \( N_K(\theta^t + 0.5 \delta \nabla \log p(\theta^t | y), \delta I_K) \) where \( I_K \) is the \( K \) dimensional identity matrix and \( \delta \) a carefully chosen variance parameter. This proposal is accepted with probability

\[
\alpha(\theta^t, \theta) = \min \left\{ 1, \frac{p(\theta | y) q(\theta, \theta^t)}{p(\theta^t | y) q(\theta^t, \theta)} \right\}
\]

where

\[
q(x, z) = (2\pi\delta)^{-K/2} \exp \left[ -\frac{1}{2\delta} \| z - x - 0.5\delta \nabla \log p(x | y) \|^2 \right]
\]

i.e. \( \theta^{t+1} \) is set equal to \( \theta \) if accepted and to \( \theta^t \) otherwise.

Roberts and Rosenthal (1998) have shown that the relative efficiency of the algorithm can be characterized by its overall acceptance rate, independently of the target distribution. The asymptotic optimal value for that last quantity is 0.57 with acceptance probabilities in the range (0.40, 0.80) still reasonable. The parameter \( \delta \) must be tuned to have an acceptance rate in that range. An automatic tuning of \( \delta \) targeting the optimal 0.57 rate is possible (Atchadé and Rosenthal, 2005).

4.2 Reparametrization of the posterior

If the tuned Langevin algorithm is directly used on the above derived posteriors, then one will observe large auto- and cross-correlations in the so-generated chain. This is not surprising as one expects that the B-splines parameters associated with neighbouring knots will take similar values, as imposed by the smoothness prior and also probably by the observed data (Lambert and Eilers, 2005). Therefore, a safe strategy consists in reparametrising the posterior before running the McMC algorithm. This can be achieved by a rough estimation of the B-splines parameters using, for example, a frequentist method for fixed and arbitrary values of the roughness penalty parameters. The IWLS algorithm is a possible choice as it quickly
provides the MLEs and the hessian of the parameters in a polytomous logistic regression model. For example, in the 1D case, we iteratively apply

$$\phi_{t+1} = (B'W_t B + \tau P)^{-1}B'(y - n\pi_t + W_t B\phi_t)$$

where $$\pi_t = \pi(\phi_t)$$ and $$W_t = \text{diag}(n\pi_t(1 - \pi_t))$$. The value of $$\tau$$ can be selected using cross-validation or an information criterion like the AIC (Eilers and Marx, 1996). The so-obtained MLE $$\hat{\phi}_t$$ and its asymptotic variance-covariance matrix $$V_\tau$$ suggest reparametrising the posterior using $$\phi'$$ where

$$\phi = V_\tau^{1/2} \phi' + \hat{\phi}_\tau.$$  

This device considerably reduces the posterior correlation and, hence, the poor mixing of the chain in the original parametrisation.

5 Illustration

5.1 Bivariate smoothing

The data of interest are the waiting times between and durations of 272 eruptions of the Old Faithful geyser. The waiting time and the duration axes were both divided into 50 bins on (35,105) and (1,6) respectively. Twenty equidistant knots were used on both axes. Two different roughness penalty parameters were allowed, one for each axis. A graphical representation of the fitted density is shown in Figure 1. The top left part shows the scatterplot together with the contours of the fitted density, rescaled to be 1 at its maximum; the top right panel displays the fitted bivariate density. The corresponding marginal densities are shown in the lower panels.

5.2 Quantile regression

A nice application of density estimation is quantile regression. Indeed, if an estimate of the joint density $$f_{Y_1 Y_2}$$ is available, then it is straightforward to derive the estimates of the corresponding marginal and conditional densities from their respective definitions.

In addition, one can derive the conditional distribution functions $$F_{Y_1|Y_2=y_2}$$, $$F_{Y_2|Y_1=y_1}$$, by integrating the corresponding density estimate. These can be inverted to finally obtain estimates of the conditional quantile functions. As an illustration, the conditional quantiles of Duration given Waiting time were derived from the fitted bivariate density, yielding Figure 2: as expected, these are smooth and do not cross as sometimes happens with some nonparametric methods. Pointwise credible interval for these deciles can also be obtained from the generated chain.
FIGURE 1. Fitted bivariate density (rescaled to be 1 at its maximum) and its marginal densities.
FIGURE 2. Waiting times between and durations of 272 eruptions of the Old Faithful geyser: conditional deciles of Duration for given Waiting time.
References


Applications of profile confidence intervals for contingency table parameters

Joseph B. Lang

Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242, USA, joseph-lang@uiowa.edu

Abstract: This work is motivated by a spate of literature that suggests that profile score and profile likelihood confidence intervals generally have better coverage properties than their Wald counterpart. It is shown by example that profile interval estimation for contingency table parameters is more broadly applicable than the current literature would suggest. The estimation approach draws on the theory of multinomial-Poisson homogeneous (MPH) models. The MPH approach lends itself to a simple computational algorithm and leads naturally to practically useful results on estimability and sampling plan invariance.

Keywords: Categorical data; multinomial-Poisson homogeneous models; profile score interval; profile likelihood interval.

1 Introduction

We describe a profile interval estimation approach for contingency table parameters that is based on the theory of multinomial-Poisson homogeneous (MPH) models (Lang 2004, 2005). The MPH approach lends itself to a simple computational algorithm that is broadly applicable. The method also leads naturally to theoretical results on estimability and sampling plan invariance. Several simple and commonly-encountered examples illustrate the computational and theoretical results.

This work is motivated by a spate of literature that suggests that profile score and profile likelihood confidence intervals generally have better coverage properties than their Wald counterpart—see for example, Meeker and Escobar (1995), Newcombe (1998a,b), Agresti and Coull (1998), Agresti and Caffo (2000), Brown et al. (2001), Tsimikas et al. (2002). Profile intervals have been used on a case-by-case basis for several different contingency table parameters, e.g., odds ratio, relative risk, and risk difference. These examples in the literature use a common computational approach that has two main limitations: it is case-specific and it is applicable for a restrictive class of parameters. The method used in this paper avoids these limitations.
2 Examples

Let $y$ and $\tau$ represent the vector of contingency table counts and probabilities and let $S(\tau)$ represent the estimand (i.e. parameter) of interest. In practice, there are two main types of estimands: Likelihood explicit estimands afford an explicit reparameterization of the data likelihood $L(\tau \mid y) \rightarrow L(S(\tau), \eta \mid y)$, where $\eta$ is a nuisance parameter. Likelihood implicit estimands afford no such explicit reparameterization. Owing to computational simplicity, most of the profile interval examples in the contingency table literature restrict attention to likelihood explicit estimands. This paper shows by example that this restriction is unnecessary.

Sections 2.1-2.3 give three simple, but illustrative, examples of estimands. Example 1 considers a likelihood explicit estimand that has been considered in the profile literature. Example 2 considers a likelihood implicit estimand that has not yet been considered in the profile literature. Example 3 describes a setting where estimability and sampling plan invariance results are needed.

2.1 Example 1. Difference Between Probabilities

Twenty-four (24) patients were sampled from Clinic 1 and 8 patients were accrued from Clinic 2 over the course of three days. Receptionists recorded whether or not each patient was insured. The data are tabled below:

<table>
<thead>
<tr>
<th>H(ealth Insurance)</th>
<th>Clinic 1</th>
<th>Clinic 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1=Uninsured</td>
<td>9</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>2=Insured</td>
<td>15</td>
<td>8</td>
<td>23</td>
</tr>
<tr>
<td>Total</td>
<td>24</td>
<td>8</td>
<td>32</td>
</tr>
</tbody>
</table>

The counts are $y = (9, 15 \parallel 0, 8)$ and the table probabilities are $\tau = (\tau_{11}, \tau_{12} \parallel \tau_{21}, \tau_{22})$. Note that $\tau_{ij} = P(H = j \mid C = i)$ so that $\tau_{1+} = \tau_{2+} = 1.0$. The estimand of interest is $D(\tau) \equiv \tau_{11} - \tau_{21}$, which is the difference between “uninsured” probabilities. It is straightforward to see that $D(\tau)$ is likelihood explicit.

2.2 Example 2. Kappa Coefficient of Agreement

Two pathologists rated 20 neonatal radiographs according to disease suspicion. The data are tabled below:

<table>
<thead>
<tr>
<th>Pathologist B Rating</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pathologist A Rating</td>
<td>10</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>
The counts are $y = (10, 1, 0 \parallel 0, 5, 0 \parallel 0, 1, 3)$ and the table probabilities are $\tau = (\tau_{11}, \tau_{12}, \tau_{13} \parallel \tau_{21}, \tau_{22}, \tau_{23} \parallel \tau_{31}, \tau_{32}, \tau_{33})$. Notice that $\tau_{ij} = P(A = i, B = j)$ so $\tau_{++} = 1$. The estimand of interest is the kappa coefficient

$$\kappa(\tau) \equiv \frac{\sum_{i=1}^{3} \tau_{ii} - \sum_{i=1}^{3} \tau_{i+} \tau_{+i}}{1 - \sum_{i=1}^{3} \tau_{i+} \tau_{+i}}.$$  

Notice that $\kappa(\tau)$ is likelihood implicit.

### 2.3 Example 3. Correlation and Gamma

A sample of 32 patients with tuberculosis (TB) was collected in Buenos Aires, Argentina, between May and August 2004. Twenty-three (23) of the 32 had uncomplicated TB, the remaining 9 had multi-drug resistant TB. Researchers were interested in determining the relationship between number of relapses and TB type (see the table of counts below).

<table>
<thead>
<tr>
<th>Relapses</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>TB Type</td>
<td>Uncomplicated = 0</td>
<td>17</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Multi-Drug Resistant = 1</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Let $\pi = (\pi_{00}, \ldots, \pi_{06} \parallel \pi_{10}, \ldots, \pi_{16})$ be the joint probabilities. For example, $\pi_{12} = P(\text{TB Type} = 1, \text{Relapses} = 2)$. The table counts are $y = (17, 3, 2, \ldots, 0 \parallel 4, 1, \ldots, 1)$ and the table probabilities are $\tau = (\tau_{00}, \ldots, \tau_{06} \parallel \tau_{10}, \ldots, \tau_{16})$. Because a row-stratified sampling plan was used, the table probabilities are conditional probabilities. For example, $\tau_{12} = \pi_{12}/\pi_{1+} = P(\text{Relapses} = 2 | \text{TB Type} = 1)$.

Among other things, it is of interest to estimate the Pearson correlation estimand $C(\pi)$ and the gamma association estimand $G(\pi)$.

### 3 Questions to be Addressed

In Example 1, one might consider using a Wald interval for estimating the estimand $D(\tau)$. The Wald interval, however, is known to have poor coverage properties when the sample sizes are small and/or the estimands fall close to a boundary. Profile confidence intervals are an attractive alternative in this setting. Currently available computational methods are estimand specific in that the log likelihood must be explicitly reparameterized in terms of the estimand, and any nuisance parameters. *Can we use a computational method that is more generally applicable, a method that does not require an explicit reparametrization of the log likelihood?* Below, we argue that the answer is yes—we describe the generally applicable multinomial-Poisson homogeneous (MPH) method.
The kappa estimand κ(τ) of Example 2 is qualitatively different than the estimand of Example 1. Specifically, κ(τ) is likelihood implicit, which means the available profile interval computational methods are not applicable (they require an explicit reparameterization of the likelihood). Can we develop a computational method that is applicable in this likelihood implicit setting? Below, we argue that the answer is yes—indeed, the MPH method used to compute profile intervals for likelihood explicit D(τ) can also be used to compute profile intervals for likelihood implicit κ(τ).

Example 3 brings up questions of estimability and sampling plan invariance. We give general results that can be used to answer the following questions: Which, if any, of the estimands C(π) and G(π) can be estimated using data y? For convenience, can we assume full-multinomial sampling when computing the profile confidence intervals for the estimable estimands? Below, we argue that under row-stratified sampling, C(π) is NOT estimable, but G(π) is estimable. We also argue that G(π) can be estimated assuming full-multinomial sampling for convenience.

4 Computing Profile Confidence Intervals

4.1 Test Inversion Intervals

In the contingency table setting, we can compute a confidence interval for estimand S(τ) by inverting the Wald, score, or likelihood ratio test of H_Δ : S(τ) = Δ versus the unrestricted alternative K_Δ : S(τ) \neq Δ. Let W^2(y, Δ), X^2(y, Δ) and G^2(y, Δ), represent the corresponding test statistics. It should be kept in mind that, unlike the score X^2 and likelihood-ratio G^2 statistics, the Wald statistic W^2 depends on the choice of parameterization.

Under non-restrictive regularity conditions on the estimand function S, these three test statistics have \( \chi^2(1) \) null limiting distributions. This implies that

\[
W_{CI} = \{ \Delta : W^2(y, \Delta) \leq \chi^2_{\alpha}(1) \}, \quad P_{SCI} = \{ \Delta : X^2(y, \Delta) \leq \chi^2_{\alpha}(1) \}, \\
P_{LCI} = \{ \Delta : G^2(y, \Delta) \leq \chi^2_{\alpha}(1) \}
\]

(1)

are approximate \( 1 - \alpha \) level Wald, profile score, and profile likelihood confidence intervals for S(τ).

There are two outstanding issues with the test-inversion confidence interval approach: (i) We must describe the regularity conditions on S that make the confidence interval formulas of (1) applicable. (ii) We must show how to find the restricted maximum likelihood estimate of τ under H_Δ : S(τ) = Δ. These issues are addressed using the theory underlying multinomial-Poisson homogeneous (MPH) models as described in the next subsection.

Remark 1. For likelihood-explicit estimands, the likelihood can be reparameterized and standard unconstrained maximization methods can be used
to find the ML estimate under $H_\Delta$. It is apparently for this reason that the majority of the contingency table examples in the profile confidence literature restrict attention to likelihood-explicit estimands. For likelihood-implicit estimands, a different approach, such as the one outlined below, is required.

4.2 MPH Models

A multinomial-Poisson (MP) model for counts $y$ can be represented as

$$y \sim MP(\sigma, \tau | Z, Z_F, n),$$

where $\sigma$ are the sample size parameters, $\tau$ are the table probabilities, and $(Z, Z_F, n)$ is the sampling plan triple. The matrix $Z$ indicates stratum membership, sampling constraint matrix $Z_F$ indicates which strata sample sizes are fixed a priori, and $n$ is the vector of a priori fixed sample sizes. Equivalently, because the expected count vector $\mu \equiv E(Y)$ is a one-to-one function of $(\sigma, \tau)$, we could write

$$y \sim MP(\mu | Z, Z_F, n),$$

keeping in mind that $\mu = D(Z\sigma)\tau$, where $D(\cdot)$ represents a diagonal matrix. MP models give rise to counts that comprise independent blocks of Poisson random variables and multinomial random vectors. Important special-case MP models include the full-multinomial model, product-multinomial model, and the Poisson model. The interested reader is directed to Lang (2004 or 2005) for more details.

An MP Homogeneous (MPH) model (cf. Lang 2004) is an MP model that constrains the expected counts $\mu$ in a special way. Specifically, an MPH model has the form

$$y \sim MP(\mu | Z, Z_F, n), \quad h(\mu) = 0,$$

where constraint function $h$ is sufficiently smooth, non-redundant, and $Z$-homogeneous, as defined in Lang (2004). An important special case that is used in this paper is when $h$ is 0-order $Z$ homogeneous: $h(D(Z\gamma)x) = h(x)$ for all $\gamma > 0$ and all $x > 0$.

Remark 2. Table probabilities in an MPH model can be written in terms of the joint probabilities as

$$\tau \equiv t(\pi) \equiv D^{-1}(ZZ^T\pi)\pi.$$

On a related note, because $Z^T\tau = 1$ (in words, the table probabilities sum to one within strata), we have that $t(\tau) = \tau$.

Remark 3. When $h$ is $Z$ homogeneous, $h(\mu) = 0$ if and only if $h(\tau) = 0$. That is, MPH models only impose constraints on the table probabilities $\tau$; they do not constrain the expected sample sizes $\sigma$.

Remark 4. For every function of the form $f(\tau)$, there exists another function $f_0$ defined as $f_0(\tau) \equiv f(t(\tau))$ that satisfies (i) $f_0(\tau) = f(\tau)$ and (ii) $f_0$ is 0-order $Z$ homogeneous. We say that $f_0$ is a 0-order $Z$ homogeneous version of $f$.

**Applicability Theorem:** Let $S_0(\tau) \equiv S(t(\tau))$ be a 0-order $Z$ homogeneous version of $S(\tau)$. Provided (S1) $S_0$ has continuous 2nd order derivatives and (S2) $S_0$ has non-vanishing gradient, the MP model under $H_\Delta$:
Profile intervals for contingency tables

\( S(\tau) = \Delta \) is an MPH model that can be expressed using the constraint
\[
b(\mu) \equiv S_0(\mu) - \Delta = 0.
\]

In words, the Applicability Theorem gives conditions on estimand function \( S \) that imply MPH theory is applicable. MPH theory as outlined in Lang (2004) can be used to argue that the Wald, score, and likelihood-ratio statistics have \( \chi^2(1) \) null limiting distributions. This means that the confidence interval formulas in (1) are applicable. Also, MPH computational algorithms can be used to compute the restricted ML estimate of \( \tau \) under \( H_\Delta \), without reparameterizing the likelihood. This means that the confidence intervals in (1) are simple to compute, even in the likelihood-implicit estimand setting.

**Remark 5.** Other MPH results in Lang (2004) apply as well. For example, it is straightforward to describe the approximate distributions of unrestricted ML estimators \( S(\hat{\tau}) \) and \( g(S(\hat{\tau})) \), where \( g \) is differentiable. This is useful for computing Wald intervals, which are used to start the computational algorithm for profile intervals.

### 4.3 An MPH-Based Computational Algorithm

**Step 1.** Compute a Wald interval for \( S(\tau) \). Let \( \Delta_{W,\text{low}} \) and \( \Delta_{W,\text{up}} \) be the endpoints.

**Step 2.** Compute the lower profile confidence bound \( \Delta_{\text{low}} \): Using \( \Delta_{W,\text{low}} \) as a reference point, find \( \Delta_1 \) and \( \Delta_2 \) satisfying
\[
X^2(y, \hat{\mu}(\Delta_1)) > \chi^2_{\alpha}(1) > X^2(y, \hat{\mu}(\Delta_2)).
\]
Here, \( \hat{\mu}(\Delta) \) is the ML estimate of \( \mu \) under \( H_\Delta \), which can be found using the MPH model fitting algorithm described in Lang (2004). With \( \Delta_1 \) and \( \Delta_2 \) as starting estimates, use a bisection algorithm to solve for \( \Delta_{\text{low}} \) in
\[
X^2(y, \hat{\mu}(\Delta_{\text{low}})) = \chi^2_{\alpha}(1).
\]
Each bisection iteration involves fitting an MPH model.

**Step 3.** Using the same approach as in Step 2, compute the upper confidence bound \( \Delta_{\text{up}} \) that satisfies \( X^2(y, \hat{\mu}(\Delta_{\text{up}})) = \chi^2_{\alpha}(1) \).

In Step 2 [3], it often works to choose \( \Delta_1 \) and \( \Delta_2 \) close to, but on either side of, \( \Delta_{W,\text{low}} [\Delta_{W,\text{up}}] \). A judicious choice of parameterization for the Wald interval can lead to better starting estimates of the profile interval bounds. Essentially the same algorithm works for profile likelihood confidence intervals; one need only replace \( X^2 \) by \( G^2 \) in Steps 2 and 3.

**Remark 6.** Assuming applicability, this MPH algorithm requires only counts \( y \) and the estimand function \( S \) as input. The sampling plan invariance results of the next section allow us to omit the sampling plan matrices \( Z \) and \( Z_F \) from the input.

### 5 Estimability and Sampling-Plan Invariance Results

Suppose that table counts \( y \) can be modeled using MPH model \( MP(\sigma, \tau|Z, Z_F, n) \). Recall that \( \tau = t(\pi) \equiv D^{-1}(ZZ^T \pi)\pi \).
Definition of Estimability: Estimand $S(\pi)$ is $Z$-estimable (i.e. estimable using $y$) if $S(\pi)$ is a function of the table probabilities, $t(\pi)$; i.e. $t(\pi_1) = t(\pi_2)$ implies that $S(\pi_1) = S(\pi_2)$, for all joint probabilities $\pi_1$ and $\pi_2$. In words, if there are two joint probabilities that give rise to different estimand values, but the same table probabilities, then counts from a table with these table probabilities can NOT be used to estimate the estimand. For example, joint probabilities $(2, 1||1, 2)/6$ and $(6, 3||1, 2)/12$ give rise to different $\pi_{11}$ values $(2/6$ vs. $6/12$), but they give rise to the same row-stratified table probabilities $(2, 1||1, 2)/3$. It follows that counts from the row-stratified table can NOT be used to estimate $\pi_{11}$. Note: Using the definition of estimability can be difficult in practice. The following theorem can help in this regard.

Estimability Theorem: Let $S_0(\pi) = S(\pi/1^T\pi)$ be a 0-order 1 homogeneous version of $S(\pi)$. Estimand $S(\pi)$ is $Z$-estimable if and only if $S_0$ is 0-order $Z$ homogeneous.

Definition of Estimation Object. The estimation object $eo[S(\tau), y, (Z, Z_F, n)]$ comprises (i) $S(\tau)$, (ii) $S(\hat{\tau})$ and $ase(S(\hat{\tau}))$, and (iii) WCI, PSCI, and PLCI.

Sampling Plan Invariance Theorem: Let $S_0$ be a 0-order $Z$ homogeneous version of $S(\pi)$. Estimand $S(\pi)$ is $Z$-estimable if and only if $S_0$ is 0-order $Z$ homogeneous.

Definition of Estimation Object. The estimation object $eo[S(\tau), y, (Z, Z_F, n)]$ comprises (i) $S(\tau)$, (ii) $S(\hat{\tau})$ and $ase(S(\hat{\tau}))$, and (iii) WCI, PSCI, and PLCI.

Corollary: If $S_0$ is 0-order $Z$ homogeneous then $S_0$ is 0-order 1 homogeneous. Thus, $eo[S(\tau), y, (Z, Z_F, n)] = eo[S_0(\tau^*), y, (Z^*, Z_F^*, n^*)]$. Here, $\tau^* = t^*(\pi) \equiv D^{-1}(Z^* Z_F^T \pi) \pi$ are table probabilities for sampling plan $(Z^*, Z_F^*, n^*)$.

In words, when $S_0$ is 0-order $Z$ homogenous and satisfies (S1) and (S2), the estimation object can be computed using the Poisson or full-multinomial model for convenience.

6 Examples Revisited

6.1 Example 1 (revisited).

The ML estimate of estimand $D(\tau)$ is $9/24 - 0/8 = 0.375$; the observed Wald 95% interval is [0.181, 0.569]. Using the MPH-based computational algorithm described in Section 4.3, we find that the profile likelihood 95% interval is [0.117, 0.574] and the profile score 95% interval is [0.021, 0.573]. Because the setting is simple and the estimand is likelihood explicit, it is not surprising that researchers have suggested many other interval estimation approaches for $D(\tau)$. As an example, the “add successes and failures” modified Wald 95% interval suggested in Agresti and Caffo (2000) is [0.021, 0.548]. Not surprisingly, this interval is numerically similar to the profile score interval.
6.2 Example 2 (revisited).

The ML estimate of $\kappa(\tau)$ is 0.835. The Wald 95% confidence interval is [0.625, 1.046]. Note: Because $\kappa$ cannot be larger than 1.0, the truncated Wald interval [0.625, 1.000] would be reported (see SAS PROC FREQ, for example.) Even though an explicit reparameterization of the likelihood in terms of $\kappa(\tau)$ is not available, the MPH-based computational algorithm of Section 4.3 is directly applicable. We find that the profile likelihood 95% interval is [0.569, 0.971] and the profile score 95% interval is [0.539, 0.954]. In contrast to the Wald interval, the profile intervals are asymmetric about the point estimate 0.835–this reflects the asymmetry in the sampling distribution of the ML estimator. Also in contrast to the Wald interval, the profile intervals do not include out-of-range values. Indeed, profile intervals have the range-preserving property.

6.3 Example 3 (revisited).

The Pearson correlation estimand has the form

$$C(\pi) = \sum_{i=0}^{1} \sum_{j=0}^{6} ij\pi_{ij} - \sum_{i=0}^{1} i\pi_{i+} \sum_{j=0}^{6} j\pi_{+j} \sqrt{\left(\sum_{i=0}^{1} i^2\pi_{i+} - \left[\sum_{i=0}^{1} i\pi_{i+}\right]^2\right) \left(\sum_{j=0}^{6} j^2\pi_{+j} - \left[\sum_{j=0}^{6} j\pi_{+j}\right]^2\right)}.$$ 

The gamma estimand has the form

$$G(\pi) = \sum_{i=0}^{1} \sum_{j=0}^{6} \pi_{ij} \left(\sum_{h>i} \sum_{k>j} \pi_{hk}\right) - \sum_{i=0}^{1} \sum_{j=0}^{6} \pi_{ij} \left(\sum_{h>i} \sum_{k<j} \pi_{hk}\right) \sum_{i=0}^{1} \sum_{j=0}^{6} \pi_{ij} \left(\sum_{h>i} \sum_{k<j} \pi_{hk}\right).$$

For estimating, we have counts $y$ from $MP(\mu|Z, Z_F, n)$, where

$$Z = Z_F = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}^T, \; n = (23, 9).$$

Note that $D(Z\gamma)x$ has form $(\gamma_1 x_0, \ldots, \gamma_1 x_6, \gamma_2 x_{10}, \ldots, \gamma_2 x_{16})^T$.

Consider the 0-order 1 homogeneous versions of the estimands, viz. $C_0(\pi) \equiv C(\pi/1^T\pi)$ and $G_0(\pi) = G(\pi/1^T\pi)$. By the Estimability Theorem, $C(\pi)$ is estimable if and only if $C_0(D(Z\gamma)x) = C_0(x)$, for all $\gamma > 0$ and $x > 0$. A quick numerical check indicates that $C_0(D(Z\gamma)x) = 0.315 \neq 0.5 = C_0(x)$ when $\gamma = (10, 1)$ and $x = (7, 6, 5, 4, 3, 2, 1)|1, 2, 3, 4, 5, 6, 7)$. We conclude that the Pearson correlation $C(\pi)$ is NOT estimable with the counts $y$.

In contrast, it is relatively straightforward to see that $G_0(D(Z\gamma)x)$ does in fact equal $G_0(x)$ for all $\gamma > 0$ and $x > 0$. We conclude that the gamma estimand $G(\pi)$ is estimable with the counts $y$. As an interesting aside, we point out that variants of gamma, such as Somer’s $d$, are NOT estimable.
Also, gamma would not be estimable if there were more than two rows in the table.

Because $G_0$ is 0-order $Z$ homogenous, the corollary to the Sampling Plan Invariance Theorem implies that we can use full-multinomial or full-Poisson sampling for convenience for computing the Wald and profile confidence intervals. The ML estimate of $G(\pi) (= G(\hat{\tau}))$ is $G(\hat{\tau}) = 0.557$. The 95% confidence intervals are as follows: Wald $[0.082, 1.033]$, Profile Score $[-0.042, 0.859]$, and Profile Likelihood $[-0.071, 0.872]$.

References


Product partition models and clustering

Vicki Livingstone\textsuperscript{1} and Claire Jordan\textsuperscript{2}

\textsuperscript{1} Department of Preventive and Social Medicine, University of Otago, P.O. Box 913, Dunedin, New Zealand
\textsuperscript{2} Institute of Information and Mathematical Sciences, Massey University, Private Bag, 102 904 NSMC, Auckland, New Zealand

Abstract: Modelling using mixture models has become increasingly popular in recent times, mostly due to their flexible nature. Mixture models have two main purposes. One is to model unknown distributional shapes and the other is to provide a model-based clustering. In this paper we focus on the latter. A mixture model is defined by a number of densities known as component densities. The number of components is typically fixed but unknown and has to be inferred from the data. Determining the true number of components in a mixture is very important and extremely difficult and has yet to be resolved. We propose using product partition models (PPMs) to determine the mixture component parameters and the true number of components in the data. PPMs allow us to partition a set of objects into \( k \) components such that objects in the same component are exchangeable and the objects belonging to distinct components are independent. They provide a convenient way of allowing the data to weight the partitions likely to hold. Markov chain Monte Carlo (MCMC) techniques are used to generate partitions of the data.

Keywords: Mixture models; clustering; product partition models; Markov chain Monte Carlo.

1 Introduction

Clustering problems arise in many areas of science. Examples include pattern recognition, data mining, machine learning and image processing in the field of computer science and the classification of plants and animals in a biology setting. A clustering problem is one where we wish to divide a set of \( n \) objects, \( X_1, X_2, \ldots, X_n \), into \( k \) non-overlapping homogeneous subsets. Each subset is referred to as a component and each object should belong to one component only such that objects within each component are “similar”.

The use of mixture models to deal with clustering problems has become popular. In a mixture model, each cluster is represented by a density, known as a component density and the entire data set is modelled by a mixture of these densities. This model-based approach attempts to optimise
the fit between the data and the model using a fixed number of components, \(k\). Determining \(k\) is not a trivial problem, and many approaches have been suggested in the literature. These include the use of reversible jump MCMC (Richardson and Green, 1997, Gruet et al 1999) and birth-and-death processes (Stephens, 2000, Cappé et al, 2002). In this paper, we propose the use of product partition models (PPMs) to estimate both the mixture component parameters and the number of components from the data. PPMs have a range of applications. Jordan and Livingstone (2005) give some examples of where they are useful. The main feature of a PPM is that the likelihood, prior and posterior for a random partition of observed data are products over partition components (Dahl, 2003). PPMs specify prior probabilities for a random partition and update these into posterior distributions of the same form. They provide a convenient way of allowing the data to weight the partitions likely to hold. A mixture model can be represented by a PPM. In this paper, we focus on univariate normal mixtures. However, our method can be applied to any set of component densities.

2 Product Partition Models

PPMs developed by Hartigan (1990) and Barry and Hartigan (1992) are used to partition a set of \(n\) objects into \(k\) components. PPMs assume that a set of objects can be partitioned into components such that each object is a member of one component only, objects in the same component are exchangeable and objects in different components of a partition are independent.

The set of all objects is defined to be \(S_0 = 1, 2, \ldots, n\). The objects in \(S_0\) are partitioned into components and a partition comprising \(k\) components is defined as \(\rho = \{S_1, S_2, \ldots, S_k\}\), such that \(S_i \cap S_j = \emptyset\) for \(i \neq j\) and \(\bigcup_i S_i = S_0\). Each possible component is assigned a cohesion \(c(S) \geq 0\) such that the prior probability of \(\rho\) is proportional to the product of the cohesions across the components comprising the partition.

For each object \(i\), there is an observation \(X_i\). Let \(X_S\) denote the vector of observations \(X_i, i \in S\) and let \(p_S(X_S)\) be the conditional density for the observations in a component \(S\), given \(S \in \rho\).

The posterior distribution of \(\rho\) given the observations is:

\[
P(\rho = \{S_1, S_2, \ldots, S_k\} \mid X) \propto P(X \mid \rho = \{S_1, S_2, \ldots, S_k\}) \times P(\rho = \{S_1, S_2, \ldots, S_k\})
\]

\[
\propto \prod_{r=1}^k c(S_r)p_{S_r}(X_{S_r}).
\]
To specify a product partition model, we need to choose the prior cohesions $c(S)$ and the prior densities $f_S(\theta^S)$ for each component $S$. We also need to specify the distribution of the data.

2.1 Model Specification

(a) Distribution of the Data.
Objects partitioned into the same component have a common mean. We let $\mu^S$ be the common mean for the $\mu_i$'s with $\mu_i = \mu_S, i \in S$.

\[
X_1 | i \in S, \mu_S, \sigma^2_S \sim N(\mu_S, \sigma^2_S); \quad i = 1, \ldots, n. \tag{1}
\]

(b) Prior Densities for the Components.
We assume that $\mu_S$ and $\sigma^2_S$ are independent with a normal and an inverse gamma prior respectively

\[
f_S(\mu_S) \sim N(\mu_0, \omega \sigma^2_S); \tag{2}
\]

where $\omega$ and $\mu_0$ are given subjective values. The precision follows a gamma distribution such that

\[
\sigma^{-2}_S \sim \Gamma(\alpha, \beta); \tag{3}
\]

where $\alpha$ and $\beta$ are chosen subjectively.

(c) Prior Cohesions.
The prior probability of a partition is proportional to a product of cohesions, one for each component of the partition. The cohesions may be used to give higher prior probability to partitions consisting of components that are considered likely a priori or of components that are, in some sense, "desirable". For example, it may be desirable to choose a prior cohesion that gives a low prior probability to partitions with many components.

There is a connection between PPMs and the Mixture of Dirichlet Processes models. Choosing cohesions of the form $C(S) = (|S| - 1)! c$ where $S_r$ is the number of objects in component $S_r$ and $c$ is a known constant results in a PPM that is a reparameterisation of a Bayesian non-parametric model with a Dirichlet Process prior (Quintana, 2003). We will consider a number of different prior cohesions and assess the sensitivity of the results to the choice of prior cohesions.

Using (1), (2) and (3) the joint posterior distribution is given by

\[
P(\rho, X, \mu_S, \sigma^2_S) \propto \left[ \prod_{r=1}^{k} C(S_r) \right] \left[ \frac{n^\alpha}{\Gamma(\alpha + \beta)} \right] \left[ \prod_{S=1}^{k} \left( \frac{1}{\omega n_S + 1} \right)^{\frac{1}{2}} \right] \times \prod_{S=1}^{k} \left[ \Gamma \left( \frac{n_S}{2} + \alpha \right) \left( \beta + \frac{n_S}{2} + \frac{n_S \mu^2 - \omega_n \sigma^2 - 2 \mu t_S}{2(\omega n_S + 1)^{\frac{3}{2}}} \right)^{-\left( \frac{n_S}{2} + \alpha \right)} \right]; \tag{4}
\]

where $v_S = \sum_{i \in S} X_i^2$, $t_S = \sum_{i \in S} X_i$ and $n_S$ is the number of objects in component $S$. 
3 The Algorithm for Generating Partitions

Our method involves determining partitions of the data. The number of possible partitions of the data increases exponentially with the number of components. It is not feasible to generate all possible partitions so Markov chain Monte Carlo methods are used to generate a sequence of partitions $\rho_0, \rho_1, \rho_2, \ldots$. The algorithm that is used is as follows:

1. Set $\rho_0 = \{1, 2, \ldots, n\}$.

2. For $i = 1, \ldots, n$:
   - Calculate the probabilities of the partitions generated by the following moves:
     - Leaving object $i$ where it is.
     - Moving object $i$ to one of the existing components.
     - Creating a new component with object $i$ by itself.
   - The probability of a partition is proportional to the joint distribution given in (4). The new partition is selected by generating a U(0,1) random number.

3. Given the partition, we calculate the expected values of the component parameters $E(\mu_i \mid X, \rho)$ and $E(\sigma_i^2 \mid X, \rho)$.

4. Repeat steps (1)-(3) $N$ times. As the partitions generated depend on the ordering of the data $X$, we randomise the order of the data at each iteration.

The posterior mean and variance for each object is a weighted average of the $\mu$ and $\sigma^2$ values associated with the components that the object was assigned to using the posterior probabilities as weights. $E(\mu_i \mid X)$ is estimated by

$$\hat{\mu}_i = \frac{1}{N - N_{bi}} \sum_{j=N_{bi}+1}^{N} E(\mu_i \mid X, \rho_j);$$

and $E(\sigma_i^2 \mid X)$ is estimated by

$$\hat{\sigma}_i^2 = \frac{1}{N - N_{bi}} \sum_{j=N_{bi}+1}^{N} E(\sigma_i^2 \mid X, \rho_j);$$

where $N_{bi}$ is the number of samples discarded during burn-in.

We also consider the posterior distribution of the number of sets in the partitions generated to determine the true number of components in the data.
4 Application to Three Real Data Sets

We demonstrate the performance of our proposed method using the so-called acidity, enzyme and galaxy data sets, which were analysed in Richardson and Green (1997). We present the results for each example, together with a description of the sensitivity of those results to the model specification. We also compare our results to those obtained by Richardson and Green.

References


Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. *Annals*
V. Livingstone and C. Jordan  

of Statistics, 28, 40-74.
Estimating the probability distribution of gamblers by observing betting practices

Caroline Lynch\textsuperscript{1}, Daniel Barry\textsuperscript{2}

\textsuperscript{1} Department of Mathematics & Statistics, University of Limerick, Ireland
\textsuperscript{2} Vice President Academic & Registrar, University of Limerick, Ireland

Abstract: A bookie seeks to maximise profit by manipulation of odds, thus influencing customers’ betting behaviour. Assuming that how customers bet depends on the probabilities they assign to the outcomes of the contest, the bookie needs information about the distribution of these probabilities. We examine how to estimate this distribution, based on observation of customers’ betting practices, in the case of a contest with two possible outcomes.

Keywords: EM algorithm; bookmaker; horse race; Markov decision process.

1 Introduction

A bookie takes bets on a contest with two possible outcomes, $A$ and $B$. Let $p$ denote a gambler’s probability that outcome $A$ will occur. The bookie quotes odds of $O_1(O_2)$ against outcome $A(B)$. Thus a winning wager of 1 unit on outcome $A(B)$ produces a return of $O_1 + 1(O_2 + 1)$. A wager on outcome $A$ will be attractive to the gambler if $p(O_1 + 1) \geq 1$ i.e., $p \geq \frac{1}{O_1 + 1} = \theta_1$. Similarly a wager on outcome $B$ will be attractive to the gambler if $1 - p \geq \frac{1}{O_2 + 1} = \theta_2$ i.e. $p \leq 1 - \theta_2$. $\theta_1(\theta_2)$ is the bookie’s “quoted probability” for outcome $A(B)$. Hence, the strategy for an individual gambler is simple - he places a wager on any outcome for which his probability exceeds that quoted by the bookie.

We assume that the bookie sells tickets which guarantee a return of 1 unit should outcome $A(B)$ occur and cost $\theta_1(\theta_2)$. We also assume that the customers’ probabilities $p_i \{i \in [1, N]\}$ of outcome $A$ occurring behave like a random sample from a probability distribution. Finally, we assume that customers can buy at most one of each type of ticket and that the bookie is free to alter the quoted probabilities after each customer.

1.1 Form of $F$

We subdivide the interval $[0,1]$ into $r$ subintervals of equal width. We then estimate $F$ by means of a histogram, with $r$ intervals - the optimal value of $r$ is found by inspection. For each interval, the height of the histogram
will be determined by the probability assigned to that interval, \( \pi_j \). This probability will be determined by the betting behaviour of the customers, as described hereafter. \( F(\theta) \) may then be determined by the piecewise formula

\[
F(\theta) = \sum_{i=1}^{j} \pi_i + (r\theta - j)\pi_{j+1}, \quad \frac{j}{r} \leq \theta \leq \frac{j+1}{r}
\]

for interval \( j \), where \( j \in [0, r] \).

2 Estimation

Each customer’s betting pattern gives us information about their value of \( p \):

- Bet on Horse A \( \theta_1 \leq p \leq 1 \)
- Bet on Horse B \( 0 \leq p \leq 1 - \theta_2 \)
- No Bet \( 1 - \theta_2 < p < \theta_1 \)

We denote the range within which \( p \) falls by \([a_1^k, a_2^k] \) and the subintervals of \([0,1] \), \( I_j \), by \([L_j, R_j] \), with \( R_j = L_{j+1} \). The range for \( p \) may, or may not, coincide with the subintervals. We have an indicator function, \( X_{jk} \), defined as \( X_{jk} = 1, p \in [L_j, R_j] \) and 0 otherwise. The log likelihood function is then \( \ell = \sum_{k=1}^{N} \sum_{j=1}^{r} X_{jk} \log \pi_j \). Given the customer’s behaviour, we have a range for the customer’s probability - i.e. \( a_1^k \leq p \leq a_2^k \). Let us call this information \( Y_k \). We seek to maximise

\[
E(\ell|Y) = \sum_{k=1}^{N} \sum_{j=1}^{r} E(X_{jk}|Y_k) \log \pi_j.
\]

\[
E(X_{jk}|Y_k) = P(X_{jk} = 1|Y_k) = \frac{P(p \in [L_j, R_j]|p \in [a_1^k, a_2^k])}{P(p \in [a_1^k, a_2^k])} = \frac{P(p \in [L_j, R_j] \cap [a_1^k, a_2^k])}{P(p \in [a_1^k, a_2^k])} = \frac{P(p \in [L_j, R_j] \cap [a_1^k, a_2^k])}{\sum_{i=1}^{r} P(p \in [L_i, R_i] \cap [a_1^k, a_2^k])}
\]

We have

\[
P(p \in [L_j, R_j] \cap [a_1^k, a_2^k]) = \pi_j \times \frac{l_{jk}}{R_j - L_j}
\]
The probability distribution of gamblers

where \( l_{jk} \) is the length of \([L_j, R_j] \cap [a^1_k, a^2_k] \) and is given by

\[
l_{jk} = \begin{cases} 
0 & R_j \leq a^1_k \\
R_j - a^1_k & L_j \leq a^1_k \leq R_j \leq a^2_k \\
R_j - L_j & a^1_k \leq L_j \leq R_j \leq a^2_k \\
a^2_k - L_j & a^1_k \leq L_j \leq a^2_k \leq R_j \\
0 & L_j \geq a^2_k 
\end{cases}
\]

3 Calculation of \( F \)

The Maximum Likelihood Estimate for \( \pi_j \) is given by

\[
\hat{\pi}_j = \frac{\sum_{k=1}^{N} E(X_{jk} | Y_k)}{N}.
\]

Each of the subintervals of \([0,1]\) was assigned an initial probability, \( \pi^1_j \). For simplicity, this initial probability was the same for each subinterval, assuming the Uniform distribution, so that, with \( r \) subintervals, the initial values of \( \pi^1_j \) are given by

\[
\sum_{j=1}^{r} \pi^1_j = 1 \Rightarrow \pi^1_j = \frac{1}{r}, \forall j.
\]

This initial probability was then updated by observing each customer’s behaviour.

3.1 Early Customers

For the first few customers, the odds are chosen so as to maximise the information obtained. We derive the information matrix, \( I \), using the formula

\[
I_{ij} = E\left[-\frac{\partial^2 \ell}{\partial \pi_i \partial \pi_j}\right].
\]

As before, we may express the log likelihood function as

\[
\ell = \sum_{k=1}^{N} \sum_{j=1}^{r} X_{jk} \log \pi_j \\
= \sum_{k=1}^{N} \left[ \sum_{j=1}^{r-1} X_{jk} \log \pi_j + X_{rk} \log \left(1 - \sum_{j=1}^{r-1} \pi_j\right)\right] \\
= \sum_{j=1}^{r-1} \left( \sum_{k=1}^{N} X_{jk} \right) \log \pi_j + \left( \sum_{k=1}^{N} X_{rk} \right) \log \left(1 - \sum_{j=1}^{r-1} \pi_j\right)
\]

since \( \pi_r = 1 - \sum_{j=1}^{r-1} \pi_j \).
Hence, we find that

\[ \frac{\partial \ell}{\partial \pi_j} = \sum_{k=1}^{N} \left[ \frac{X_{jk}}{\pi_j} - \frac{X_{rk}}{1 - \sum_{j=1}^{r-1} \pi_j} \right] \]

and

\[ \frac{\partial^2 \ell}{\partial \pi_i \partial \pi_j} = \frac{\sum_{k=1}^{N} X_{jk} \delta_{ij} - \sum_{k=1}^{N} X_{rk}}{(1 - \sum_{j=1}^{r-1} \pi_j)^2} \]

Thus, we have

\[ E \left[ -\frac{\partial^2 \ell}{\partial \pi_i \partial \pi_j} \right] = \frac{N \delta_{ij}}{\pi_j} + \frac{N}{1 - \sum_{j=1}^{r-1} \pi_j} \]

The entries are added for each successive customer.

We now choose the odds for the initial customers. Firstly, both \( \theta_1 \) and \( \theta_2 \) are set at \( \frac{1}{r} \). The information matrix is recalculated for each combination of \( \theta_1 \) and \( \theta_2 \), each being incremented in steps of \( \frac{1}{r} \). That combination of odds which maximises the determinant of the information matrix is used for the next customer. This procedure is repeated for each of the customers in turn. The optimal number of customers used to estimate \( F \) is found by inspection. After each of these customers bets, our estimate of \( F \) is updated using the method described previously.

### 3.2 Odds for the First Customers

To initialise the information matrix, we follow the following plan:

If we divide the interval \([0,1]\) into \( r \) equally-spaced subintervals, placing the \( \theta \)-values on the divisions of these subintervals will give us precise information about the distribution of probability within these subintervals. We also know that we do not need to set either a \( \theta \)-value of 1, which guarantees no bets, or 0, which guarantees a bet from any customer.

We set the \( \theta \)-values for the first customer as \( \theta_1 = \theta_2 = \frac{r-1}{r} \). We then take each theta-value down by a value \( \frac{1}{r} \) in turn for each of the next few customers. The fact that customers bet on outcome \( A \) with probability \( 1 - F(\theta_1) \) means that the value of \( \theta_1 \) gives us information about the probabilities of the subintervals above \( \theta_1 \). Similarly, customers bet on outcome \( B \) with probability \( F(1-\theta_2) \), so the value of \( \theta_2 \) gives us information about the subintervals below \( \theta_2 \) - thus, the \( \theta \)-values for the first customer give us information about the first, and last, subintervals. Since the probabilities sum to 1, we only need information about \( r-1 \) subintervals - in all, therefore, we need \( r-2 \) customers in this first group.

### References

Spline smoothing in Bayesian disease mapping

Ying C. MacNab\textsuperscript{12} and Paul Gustafson\textsuperscript{3}

\textsuperscript{1} Department of Health Care and Epidemiology, University of British Columbia
\textsuperscript{2} Centre for Healthcare Innovation and Improvement, British Columbia Research Institute for Children’s and Women’s Health, Room E417, 4480 Oak Street, Vancouver, BC, Canada V6H 3V4
\textsuperscript{3} Department of Statistics, University of British Columbia, Canada

Abstract: In the context of Bayesian disease mapping, we present semiparametric models with spline smoothing and associated prior specifications for a random splines ensemble for Bayesian predictive inference of unobserved spatio-temporal relative risks. We present smoothing with regression B-splines, smoothing splines and P-splines and a comparison of the three approaches within the context of risks ensemble prediction. The methods are motivated and illustrated through a Bayesian analysis of iatrogenic injuries to hospital in-patients in British Columbia, Canada.

Keywords: Bayesian disease mapping; semiparametric models; regression B-spline; smoothing spline; P-spline.

1 The Bayesian disease mapping model framework

In the context of disease mapping, modelling spatio-temporal disease rates requires considerations of spatial smoothing, temporal smoothing and spatio-temporal interaction. In this paper, these considerations are motivated and explored through development of a class of semi-parametric mixed effects models with spline smoothing. The methodological development aims to provide reliable information about the patterns (both over space and time) of disease risks and to quantify uncertainty.

Let $y_{it}$ denote disease occurrence and $n_{it}$ the ‘at risk’ population for the $i$th local region at time $t$; $i = 1, ..., N$ and $t = t_1, ..., t_T$. We discuss three-stage hierarchical semiparametric models and assume that conditioning on a vector $b$ of random effects the data $\{y_{it}\}$ arise independently from Poisson family with intensity $\mu_{it}$:

$$\log(\mu_{it}|b) = \log(n_{it}) + a_0 + S_0(t) + b_{i0} + RS_i(t). \quad (1)$$

Here, $\log(n_{it})$ is an offset; $S_0(t)$ and $RS_i(t)$ represent arbitrary smoothing functions. Let $t$ be centered, $a_0 + S_0(t)$ represents ‘global’ rate trend, with $m = \exp(a_0)$ representing mid-period ‘global’ rate; $\{b_{i0} + RS_i(t)\}$ represent
regional risk trend relative to the ‘global’ rates, with $b_{i0}$ representing the mid-period random area effect.

2 Temporal smoothing with B-splines

We present three spline methods for fitting of the smoothing functions $S_0$ and $RS_i$: Regression B-spline, smoothing spline, and P-spline. The key feature of the regression spline method is that for a given set of knots, the smoothing functions are fit by mixed-effects regression in terms of the B-splines; the smoothing functions are fit by piecewise cubic polynomials with a minimal number of interior knots. As exemplified herein, the approach allows us to blend together spatial, temporal and nonlinear modeling within one unified Bayesian disease mapping framework to enhance modeling of disease and health outcomes data.

Specifically, we assume a regression cubic B-spline for the arbitrary smoothing function $S_0(t)$ and a family of regression cubic B-splines for $RS_i(t)$:

$S_0(t) = \sum_{k=1}^{K} a_k B_k(t)$ and $RS_i(t) = \sum_{k=1}^{K} b_{ik} B_k(t)$, where $\{a_k, k = 1, \ldots, K\}$ are fixed effects; $\{B_k\}$, $k = 1, 2, \ldots, K$, is a set of basis functions (without the intercept) for a $K$-dimensional space $S_0(t_i, \ldots, t_L, 3)$ of B-splines of degree 3, $(t_i, \ldots, t_L)$ denoting $L$ pre-specified inner knots, $B_k(t)$ denoting the $k$th B-spline basis function evaluated at time $t$, $K = L + 3$; $b_{ik} = (b_{1k}, \ldots, b_{Nk})^\top$ are random spline coefficients. In our disease mapping applications, we have been working with annual data of 10 years of time span. Note that in common Bayesian disease mapping application, one often models sparse data with underlying rate changes gradually over space and time. For modeling data with a relatively small number of equally spaced time points (say about 10 years), we propose a simple and computationally efficient method for knot selection as well as model selection. The method begins with placing one interior knot at mid-period and give specific attention to the assessment of temporal smoothing and goodness-of-fit for models of increasing number of inner knots, positioned uniformly over the range of time $t$. Knot and model selections are based on examining the corresponding deviance residuals, deviance, and DIC output from the MCMC implementation. In general, one would expect that the 1- or 2-knot model to smooth out sharply fluctuate extreme rates and unveil gradually changing risk patterns (both over space and time). With increasing number of interior knots, the B-splines should explore increasingly localized variation and impose lesser degree of temporal smoothing; the corresponding relative risk estimates will reflect increased uncertainty and have wider credible intervals. It should be noted that the $NK$-vector of regression parameters $b = (\beta_{i10}, \ldots, \beta_{N0}, \ldots, \beta_{iK}, \ldots, \beta_{NK})^\top$ represents a random effects ensemble that governs the spatio-temporal risks smoothing via specifications of the second-stage prior. In this paper, we consider two prior assumptions, one assumes spatially varying smoothing splines (the spatial prior) and the
other randomly varying smoothing splines (the independence prior). For the spatial prior, we consider the following conditional autoregressive (CAR) formulation for ‘area adjacency’ defined neighborhood setup[1]:

\[ b_k \sim N(0, \Sigma_k), \quad \Sigma_k = \sigma_k^2 D^{-1}, \quad D_k = \lambda_k R + (1 - \lambda_k)I_N, \quad k = 0, 1, \ldots, K. \]  

(2)

where \( \sigma_k^2 \) quantifies spatial dispersion and \( \lambda_k \) spatial autocorrelation, \( \lambda_k \in (0, 1) \); \( I_N \) is an identity matrix of dimension \( N \); \( R \) is the commonly known “neighborhood matrix” of dimension \( N \): its \( i \)th diagonal element equals to the number of neighbors for the \( i \)th local area and the off diagonal elements equal to minus one for areas that are neighbors and zero otherwise. We also consider the following unstructured multinormal prior:

\[ b_i = (\beta_{i0}, \beta_{i1}, \ldots, \beta_{iK})^\top \sim N(0, \Sigma), \quad i = 1, \ldots, N, \]  

(3)

where \( \Sigma \) is an unstructured \( K + 1 \) by \( K + 1 \) variance-covariance matrix with elements \( \sigma_{uv}, u = 1, \ldots, (K + 1), v = 1, \ldots, (K + 1) \).

3 Temporal smoothing with penalty splines:  
Smoothing spline and P-spline

In comparison with the ‘few-knot’ B-spline, the smoothing spline and P-spline are splines of many knots. To regulate goodness-of-fit and smoothness, the corresponding analyses are formulated in terms of penalized regressions [2-4]. A smoothing spline has one knot in each distinct data point and the ‘second derivative penalty’, while P-spline is usually formulated as a B-spline solution with respect to a difference penalty on the coefficients of adjacent B-splines. More recently, the penalty splines have been reinterpreted as empirical Bayes or fully Bayesian solutions to mixed-effects models with assumptions on the design matrices and on the components of variance of the random effects.

Both the smoothing spline and P-spline approaches can be readily casted into the GLMM context and the ensuing Bayesian framework for disease mapping. For example, the smoothing spline GLMM formulation of model (1) can be written as

\[ \log(\mu_{it} | \mathbf{b}) = \log(n_{it}) + u_0 + u_1 t + \sum_{k=2}^{T-1} u_k C_k(t) + \nu_{i0} + \nu_{i1} t + \sum_{k=2}^{T-1} \nu_{ik} C_k(t), \]  

(4)

where \( (u_0, u_1) \) are fixed effects, \( \mathbf{u} = (u_2, \ldots, u_{T-1})^\top \) are random effects, \( \mathbf{u} \sim N(0, \sigma_u^2 I_{T-1}) \), and \( \sigma_u^2 \) is the smoothing parameter; \( (\nu_{i0}, \nu_{i1}) \) and \( \mathbf{\nu}_i = (\nu_{i2}, \ldots, \nu_{iT-1})^\top \) are random effects, \( \mathbf{\nu}_i \sim N(0, \sigma_i^2 I_{T-1}) \), and \( \sigma_i^2 \) is the smoothing parameter, \( i = 1, \ldots, N; \quad C = (C_k(t)) \) is the corresponding \( T \) by \( T - 2 \) design matrix for the random effects, \( C = L(L^\top L)^{-1}, \) \( L = QU^{-1}, \)
U is the Choleski decomposition of R, R and Q are the two band matrices defined in [3].

Similarly, the P-spline GLMM formulation of model (1) can be written as

$$ \log(\mu_i) = \log(n_{it}) + \alpha_0 + \alpha_1 t + \sum_{k=2}^{T} \alpha_k A_k(t) + \beta_0 + \beta_1 t + \sum_{k=2}^{T} \beta_k A_k(t), $$

where \((\alpha_0, \alpha_1)\) are fixed effects, \(\alpha = (\alpha_2, ..., \alpha_T)^T\) are random effects, \(\alpha \sim N(0, \sigma_\alpha^2 I_T)\), \(\sigma_\alpha^2\) is the smoothing parameter; \((\beta_0, \beta_1)\) and \(\beta_i = (\beta_2, ..., \beta_{i+1})^T\) are random effects, \(\beta_i \sim N(0, \sigma_\beta^2 I_T)\), \(\sigma_\beta^2\) is the smoothing parameter, \(i = 1, ..., N\); \(A = (A_k(t))\) is the corresponding \(T\) by \(T - 1\) design matrix for the random effects, \(A = BD^T(DD^T)^{-1}\), \(B\) is the corresponding B-spline bases matrix, \(D\) is the difference matrix defined in [4]. Note that the P-spline method accommodates arbitrary knot placement and higher order difference penalty; for comparisons of the three smoothing methods, however, second-order difference and a set of knots at the (unique) values of the time \(t\) are assumed for the derivation of the P-spline formulation (3).

Also for the comparison purpose, we explore two ways of formulating the second-stage priors: for the smoothing splines, we consider

$$ u \sim N(0, \sigma_u^2 I_{T-2}), \nu_{00} \sim N(0, \sigma_{\nu_0}^2), \nu_{11} \sim N(0, \sigma_{\nu_1}^2), \nu_i \sim N(0, \sigma_{\nu_i}^2), \nu_{i1} \sim N(0, \sigma_{\nu_{i1}}^2) $$

or

$$ u \sim N(0, \sigma_u^2 I_{T-2}), \nu_{00} \sim N(0, \sigma_{\nu_0}^2), \nu_{11} \sim N(0, \sigma_{\nu_1}^2), \nu_i \sim N(0, \sigma_{\nu_i}^2), \nu_{i1} \sim N(0, \sigma_{\nu_{i1}}^2), i = 1, ..., N. $$

For the P-splines we assume

$$ \alpha \sim N(0, \sigma_\alpha^2 I_{T-1}), \beta_0 \sim N(0, \sigma_\beta^2), \beta_{11} \sim N(0, \sigma_\beta^2), \beta_i \sim N(0, \sigma_\beta^2 I_{T-1}) $$

or

$$ \alpha \sim N(0, \sigma_\alpha^2 I_{T-1}), \beta_0 \sim N(0, \sigma_\beta^2), \beta_{11} \sim N(0, \sigma_\beta^2), \beta_i \sim N(0, \sigma_\beta^2 I_{T-1}), i = 1, ..., N. $$

Note that for single-curve smoothing spline and P-spline models, their mixed-effects model formulations require that \(\nu_{00}s, \nu_{11}s, \beta_{00}s, \) and \(\beta_{11}s\), the coefficients with respect to the constant and linear functions, to be fixed effects. In this study, \(\nu_{00}s, \nu_{11}s, \beta_{00}s, \) and \(\beta_{11}s\) are treated as random effects to accommodate ‘pooling-data’ across small-areas. In particular, the smoothing spline (7) and P-spline (9) facilitate ‘borrowing-strength’ across small areas, both for the linear trend components (the intercepts and slopes) as well as the non-linear components (the random spline coefficients) of the smoothing functions. Note also that under smoothing spline (6), as well as P-spline (8), the relative risk splines are fit with the same degree of smoothness, as we assume a single smoothing parameter \(\sigma_\alpha^2\) for all small areas. By assuming varying smoothing parameters in (7) and (9), we allow the degree of smoothing to differ from area to area, and unequal variances for the resulting splines.
344 Spline smoothing in Bayesian disease mapping

4 The third-stage hyperpriors

We assume ‘non-informative’ and flat hyperpriors for the third-stage hierarchical specifications. Specifically, diffuse normal prior $N(0, 0.0001)$ was placed on all fixed effects and commonly used ‘non-informative’ priors were assumed for the components of variance: $\sigma^{-2} \sim \text{Gamma}(0.0001, 0.0001)$ for all variance parameters, and $\lambda \sim \text{Unif}(0, 1)$ for all correlation parameters. Under the unstructured prior (3), a vague Wishart prior was placed on the precision matrix: $\Sigma^{-1} \sim \text{Wishart}(R^{-1}, p)$, where $R$ is a $(K+1) \times (K+1)$ identity matrix, $p = K + 1$ is the shape parameter.

5 Results and conclusion

The Bayesian disease mapping models with spline smoothing were applied to the analysis of hospital admission/separation data on non-fatal iatrogenic injuries among male patients 1-19 years of age 84 local health areas in the Provence of British Columbia (BC), Canada. For the B-spline models, the DIC results and the posterior rate and risk splines favored the non-spatial prior (3) and suggested that the 4-knot model offered best balance between goodness-of-fit and smoothness (see Table 1). The analysis also suggested that the 1- and 2-knot models provided sufficient flexibility to expose gradually changing temporal trends. The 5-knot splines show clear signs of overfit, both in terms of the posterior splines and the DIC result. The hyperprior choices for the Wishart prior had some but considerably minor influence on the rate and ratio smoothing.

The smoothing spline and P-spline fittings were implemented for the aforementioned models and prior and hyperprior specifications. Overall, the two penalized smoothing methods led to similar posterior predictions of the relative risks ensemble. The smoothing spline (6) and P-spline (8) were comparable to the 1-knot B-spline (3), although the first two splines imposed visibly higher smoothing to the rates/risks. The smoothing spline (7), P-spline (9), and the 4-knot B-splines (3) were comparable smoothers; measurable differences were only occurred in a few areas that had erratic rate/ratio change or occasional extreme rates over time. The penalty splines showed comparably less smoothing and higher posterior standard deviations and wider credible intervals (i.e., higher uncertainty). Of note is that smoothing spline (7) and P-spline (9) showed considerable sensitivity with respect to hyperprior specifications of the random effects variances.

For disease mapping of 10-year period, the B-spline method offered sufficient flexibility to fit the data and compared favorably with the smoothing spline and P-spline methods. The study also showed that the smoothing splines and P-splines could be useful smoothing alternatives, particularly for data covering a longer time period.

We refer to Table 1 for a summary of our main results.
Y.C. MacNab and P. Gustafson 345

Table 1. Deviance information criterion for selected models.

<table>
<thead>
<tr>
<th>Model Description</th>
<th>Dbar</th>
<th>pD</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-knot B-spline model with prior (2)</td>
<td>4241</td>
<td>169</td>
<td>4410</td>
</tr>
<tr>
<td>1-knot B-spline model with prior (3)</td>
<td>4178</td>
<td>206</td>
<td>4384</td>
</tr>
<tr>
<td>2-knot B-spline model with prior (3)</td>
<td>4130</td>
<td>236</td>
<td>4366</td>
</tr>
<tr>
<td>3-knot B-spline model with prior (3)</td>
<td>4041</td>
<td>272</td>
<td>4314</td>
</tr>
<tr>
<td>4-knot B-spline model with prior (3)</td>
<td>4014</td>
<td>295</td>
<td>4310</td>
</tr>
<tr>
<td>5-knot B-spline model with prior (3)</td>
<td>4002</td>
<td>311</td>
<td>4313</td>
</tr>
<tr>
<td>P-spline with prior (6)</td>
<td>4233</td>
<td>182</td>
<td>4416</td>
</tr>
<tr>
<td>P-spline with prior (7)</td>
<td>4118</td>
<td>248</td>
<td>4365</td>
</tr>
<tr>
<td>Smoothing Spline with prior (8)</td>
<td>4245</td>
<td>173</td>
<td>4419</td>
</tr>
<tr>
<td>Smoothing Spline with prior (9)</td>
<td>4068</td>
<td>270</td>
<td>4337</td>
</tr>
</tbody>
</table>

Acknowledgments: We gratefully acknowledge the support of the Canadian Institute for Health Research (CIHR MOP-42498) and the Natural Sciences and Engineering Research Council of Canada (NSERC RG-238660)

References


Analysis of a cardiovascular safety experiment with longitudinal data using penalized smoothing splines

Maringwa J. T.1, Geys H.1,2, Shkedy Z.1, Faes C.1, Molenberghs G.1, Aerts M.1, Van Ammel K.2, Teisman A.2 and Bijnens L.2

1 Center for Statistics, Universiteit Hasselt, B 3590 Diepenbeek, Belgium
2 J & J Pharmaceutical Research and Development, Beerse, Belgium

Abstract: Several pharmacological studies involve experiments aimed at testing for a difference between treatment groups wherein the data are longitudinal in nature, with long sequences per subject. Imposing a parametric mean structure is often too restrictive. A more flexible approach is to model the mean using a semiparametric smooth function estimated using penalized smoothing splines. We formulate a series of semiparametric models showing how the group-specific mean profiles could possibly differ. Once an appropriate model is chosen, interest lies in identifying ‘peaks’ or specific time points where the groups differ. For this purpose, we propose the use of simultaneous confidence bands around the fitted functions wherein the bands take into account both within and between-subject variability as well as variability due to smoothing.

Keywords: Linear mixed model; penalized smoothing spline; semiparametric mixed model; simultaneous confidence bands; smoothing parameter.

1 Introduction

Testing for a difference between two treatment groups is common in many applications. We consider an example arising in the context of cardiovascular safety experiments, where a comparison between a control and a treated group is required. Data collected are of a longitudinal nature, where repeated measurements are obtained from each animal every minute over a period of 240 minutes. All measurements from each animal are available. Heart rate (beats per minute) profiles from 14 dogs in the control group and 14 dogs in the treated group are available and illustrated in Figure 1. The profiles show substantial between as well as within-subject variability. The null hypothesis that there is no treatment effect can be stated as $H_0 : f_A(t) = f_B(t)$, where $f_A(t)$ and $f_B(t)$ are the mean profiles in the two groups at time $t$. A test for a difference between the average profiles gives an overall impression about the equality of the two functions. To identify specific sections of the profiles showing significant differences between the groups, a parametric or nonparametric test can be applied at each time
point. This approach however suffers the problem of multiple comparisons especially with such long sequences and more importantly, the correlation among the observations is ignored. A comparison based on confidence bands is an attractive alternative. Here we make adjustments to the confidence bands of Ruppert et al. (2003) to accommodate the longitudinal nature of the data at hand. Specifically, the bands include a component of the variance of the random intercept, which is used to capture the correlation structure amongst the observations.

2 Model Formulation

The data fall within the framework of continuous longitudinal data and can be modeled by a linear mixed model (Verbeke and Molenberghs, 2000). We assume a random intercept model, accounting for the clustering nature
FIGURE 2. Hypothetical models illustrating how mean profiles can possibly differ

of the observations. Imposing a parametric function to model the mean evolution may be too restrictive or inadequate. A more flexible route, situated within the framework of mixed models, is to model the mean with a semiparametric smooth function, \( f(t) \), which can be estimated, among others, with penalized splines (Ruppert et al., 2003). The equivalence between the penalized spline smoother and the optimal predictor in a mixed model framework is a key feature in fitting the models. Applications of this nature can be found for example in Zeger and Diggle (1994), Verbly et al. (1999), Ruppert et al. (2003), Durbin et al. (2005).

The semiparametric model mentioned above implies that the mean response for each group can be represented by an additive model of two components: a linear and a smooth component. Figure 2 illustrates, with hypothetical examples, several possible scenarios related to the evolution of the means over time. In panel A, the two groups have the same mean hence the null model. Panel B reveals a pattern where the means of the two groups differ only by a constant while in panel C the groups are different in the linear part but the smooth component of the mean is identical. Finally, in panel D, the means of the two groups have different evolutions over time and the groups are different in both the linear and the smooth part. Each of these scenarios can be modeled in a linear mixed model framework as summarized in Table 1.

We illustrate the formulation of the most complex model (Model 5), from which other models are readily deduced. This model can conveniently be
represented by its penalized spline formulation as follows:

\[ Y_{ij} = \begin{cases} 
β_0 + β_1 t_{ij} + \sum_{k=1}^{K} b_k^A (t_{ij} - κ_k) + b_{0i} + ε_{ij}, & \text{Group A,} \\
(β_0 + β_{01}) + (β_1 + β_{11}) t_{ij} + \sum_{k=1}^{K} b_k^B (t_{ij} - κ_k) + b_{0i} + ε_{ij} & \text{Group B} 
\end{cases} \] (1)

where \( Y_{ij} \) is the response for subject \( i \) at time \( j \), \( κ_1, \ldots, κ_K \) are knots, \( b_{0i} \) is the random intercept and \( ε_{ij} \) are the error terms. Under the mixed model representation

\[ Y = Xβ + Zb + ε \] (2)

the matrix \( Z \) of random effects is block-diagonal with the diagonal elements as matrices appropriately constructed from truncated lines basis and subject-specific columns of ones and zeros, corresponding to each group. The random effects vector is defined as:

\[ b = [b_1^A, \ldots, b_K^A, b_1^B, \ldots, b_K^B, b_{01}, \ldots, b_{0n}]' \]

The covariance matrix is a block-diagonal matrix with entries for the variance components corresponding to the two groups and the random intercept given by:

\[ G = \text{Diag} [σ_{b,k}^2 I_K, σ_{b,k}^2 I_K, σ_{b,0}^2 I_n] \]

where \( σ_{b,k}^2 = \text{var}(b_k^A) \), \( σ_{b,k}^2 = \text{var}(b_k^B) \) and \( σ_{b,0}^2 = \text{var}(b_{0i}) \).

3 Testing

We consider two inferential approaches, first, based on formal hypotheses tests in Section 3.1 and then through use of confidence bands in Section 3.2.

3.1 Hypotheses Testing

Based on Akaike’s Information Criterion (Burnham and Anderson, 2002), it can be seen that a model yielding separate curves for the two groups smoothed separately with the same smoothing parameter is the best (Model 4). A formal likelihood ratio test (LRT) to see if the smoothing parameter should be varied by group yields a non-significant result (\( p = 0.7518 \)) compared to a \( χ^2_1 \). Hence we need to smooth the groups separately but with the same level of smoothing. The mean profiles in the treatment groups with overlaid fitted profiles from Model 4 are shown in Figure 3 (left panel). Selection of this model already suggests a difference between the two groups, however a formal test may be more informative. Testing Model 4 against
Model 1, the LRT statistic is highly significant \((p = 0.0001)\) when compared to a \(\chi^2_2\). Suppose Model 5 was the best model, then testing it against Model 1 involves the hypotheses formulated as:

\[
H_0 : \beta_0 = 0, \quad \beta_1 = 0, \quad \sigma^2_{\epsilon} = \sigma^2_{\beta}.
\]

For \(\sigma^2_{\epsilon} = \sigma^2_{\beta} + \Delta\), the hypothesis \(\sigma^2_{\epsilon} = \sigma^2_{\beta}\) is equivalent to \(H_0 : \Delta = 0\). The distribution of the LRT statistic is then \(\chi^2_3\) since (4) above is equivalent to:

\[
H_0 : \beta_0 = \beta_1 = \Delta = 0.
\]

Testing Model 2 or 3 against Model 1 involves classical LRT for fixed effects only. Note, the models could as well be tested in a hierarchical way, starting with Model 5 as the most complex and try to reduce it to Model 4 etc. One interesting comparison would be between Model 4 and Model 3. The two models have exactly the same number of parameters although the random effects in Model 4 are independent from function to function. A formal test for Model 4 versus Model 3 for example using a LRT is not possible since the difference in the number of parameters is 0. In such cases differentiation between the models can be based on information criteria.

### 3.2 Constructing Confidence Bands

In line with one of our primary aims, we construct 95% confidence bands based on the selected model. The main focus is to provide an adjustment of the confidence bands proposed by Ruppert et al. (2003) to account for the clustering nature of the measurements from each subject. The bands are based on simulations assuming a multivariate normal distribution for the vector of contrasts between the estimated and true parameters for both fixed and random effects.

Let \(R = \text{Cov}(\epsilon)\), \(G = \text{Cov}(\beta)\), \(B = \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} \end{bmatrix}\),

\[
D = \text{Diag}(0, 0, 1, \ldots, 1) \quad \text{and} \quad C \equiv [X \ Z].
\]

According to Ruppert et al. (2003), it can be shown that

\[
\text{Cov}
\left[
\begin{bmatrix}
\hat{\beta} - \beta \\
\hat{b} - b
\end{bmatrix}
\right]
\simeq

\left(C^T \hat{R}^{-1} C + \hat{B}\right)^{-1}.
\]

Ruppert et al. (2003) provide expressions to obtain pointwise as well as simultaneous confidence bands for independent data, which are special forms of the general form in (5).

In case of longitudinal data, one can rewrite the mixed model formulation (2) with the matrix \(Z\) split into two components as:

\[
Y = X \beta + Z^T b^b + Z^b b^b + \epsilon.
\]
FIGURE 3. Model 4: Mean and fitted profiles (left), simultaneous confidence bands (right)

where $Z^{h_0}$ corresponds to the random intercepts and $Z^b$ corresponds to smoothing. Assuming $\text{Cov}(\varepsilon) = \sigma^2 I$, it follows that

$$\text{Cov}(\varepsilon^*) = \sigma^2 b_0 J_{n_i \times n_i} + \sigma^2 \varepsilon I_{n_i \times n_i},$$

where $J$ is a matrix of ones and only ones. It can be shown that

$$\text{Cov} \left( \begin{bmatrix} \hat{\beta} - \beta \\ b - b \end{bmatrix} \right) \simeq \sigma^2 \left\{ \sum_{i=1}^{n} \left\{ C_i^T \left( I_{n_i \times n_i} - \frac{\sigma^2 b_0}{\sigma^2 + n_i \sigma^2 b_0} J_{n_i \times n_i} \right) C_i \right\} + \frac{\sigma^2}{\sigma^2 b_0} D \right\}^{-1}. \tag{6}$$

Expression (6) accommodates the correlation amongst the observations through the variance of the random intercept, $\sigma^2 b_0$. This expression can therefore be used to construct both pointwise confidence intervals and simultaneous confidence bands for Model 4.

The simultaneous confidence bands illustrated in Figure 3 (right) were found to be slightly wider (1.13 times) than the corresponding pointwise confidence bands. An approximate line indicating the position where the groups start to differ (significantly) has been drawn at time equal to 25 minutes. It is clear from Figure 3 (right panel) that the confidence bands do not overlap before time 25 minutes. This suggests that the groups only differ in the early stages of the experiment and thereafter, no significant differences are observed.
4 Discussion

The models considered here make use of the correspondence between the linear mixed model and the penalized smoothing splines. We have shown that one can formulate different possible situations illustrating how the two groups can differ. From these, a 'suitable' model is selected based on the AIC criterion. Testing for a difference in the average profiles by first fitting an overall common average curve may in certain cases involve testing for both fixed and random effects (variance components). This can formally and more appropriately be done by way of simulating the distribution of the likelihood ratio test as suggested by Crainiceanu et al. (2005) and Ruppert et al. (2003). Here we have considered the likelihood ratio tests based on asymptotic chi-square distributions. The models considered could as well be tested in a hierarchical way wherein one would attempt to reduce the most complex model (Model 5) in steps. In certain instances, it is observed that formal parametric tests between models may not be possible. In such situations, differentiation between the models can be done based on information criteria.

The detection of particular sections of the profiles showing significant differences is achieved by constructing confidence intervals and bands. We have illustrated an adaptation of the confidence intervals and bands of Ruppert et al. (2003) to accommodate the random intercept in case of longitudinal data. Extension to models including random intercepts and slopes is also possible. Pointwise confidence intervals suffer the draw backs associated with multiple comparisons wherein the overall significance level needs to be protected. To counter that, simultaneous confidence bands have been discussed specifically focusing on application to the random intercept model.

Acknowledgments: We gratefully acknowledge the financial support from the IAP research network nr. P5/24 of the Belgian Government (Belgian Science Policy).
References


Bayesian hierarchical models for hake stocks

Jorge M. Mendes$^1$ and K.F. Turkman$^2$

$^1$ Higher Institute of Statistics and Information Management, New University of Lisbon, 1070-124 Lisbon, Portugal, jmm@isegi.unl.pt
$^2$ DEIO, FCUL, University of Lisbon, Bloco C6, Campo Grande, 1749-016 Lisbon, Portugal, kfturkman@fc.ul.pt

Abstract: Two alternative Bayesian Hierarchical models are suggested to model over dispersed, zero inflated spatial count data. These models are then used to make inference on the abundance index of hake stocks in Portuguese coast.

Keywords: Over dispersed count data; MCMC; hake stocks; zero-inflation; log-Gaussian models.

1 Introduction

Portuguese bottom trawl surveys have been carried out by the Portuguese Institute for Fisheries and Sea Research (IPIMAR - Instituto de Investigação das Pescas e do Mar) on the Portuguese continental waters since June 1979, twice a year in Summer and Autumn. The gear used has been a Norwegian Campbell Trawl 1800/96 (NCT) with a codend of 20 mm mesh size, mean vertical opening of 4.8 m and mean horizontal opening between wings of 15.6 m (ICES, 2002).

The sampling design of these surveys follow a stratified random sampling design with 97 sampling locations distributed over 12 sectors. The sampling locations are fixed from year to year and were selected based on historical records of clear tow positions and the allocation of at least two samples by stratum. Each sector was subdivided into 4 depth ranges: 20-100m, 101-200m, 201-500m and 501-750 m, with a total of 48 strata. The tow duration during the period of analysis was 60 minutes.

The data we have are the counts of recruits (individuals with age 0) in hauls of 60 minutes collected by the Autumn surveys from 1990 to 1999, between 20 and 500 meters of depth. The observed data contain an excessive number of zeros as well as some large, extreme values. For example, the data set contains 99 zero counts out of 272 observations and the range, excluding the zero counts, is 606. Although, data covering the whole Portuguese continental shelf is available, this study is restricted to the data covering the northern shelf. From exploratory analysis, we conclude that there is no substantial spatio-temporal interactions in the data, which worth modelling, considering the short (10 years) temporal component. Annual data
at each spatial locations show remarkably stable, trend free behavior and therefore, we concentrate on modelling the spatial structure. In model 1 we use the whole spatial-temporal 272 observations, but because it is difficult to replicate the same sampling spatial location across time, the data have many counts that are made at longitude and latitude coordinates with a very small shift. However these are observations of the same sampling location across time. Therefore, in model 2 we averaged their counts and spatial coordinates. The final data set contains a total of 95 observations.

2 Models

Let \( Z = \{Z(s, t), s \in D \subset \mathbb{R}^2, t = 0, 1, 2, ..., 10\} \) be the observed count (hake recruitment) at year \( t \) and spatial location \( s_i = (\text{latitude}_i, \text{longitude}_i) \). The data consist of \( z(s_i, t), i = 1, 2, ..., M_t \) at \( M_t \) spatial locations, \( t = 1, 2, ..., 10 \), as well as a set of covariates \( X(s, t) = \{X_k(s, t), s \in D \subset \mathbb{R}^2, t = 0, 1, 2, ..., 10\}, k = 1, ..., K \), with observations \( x_k(s_i, t), i = 1, 2, ..., M_t \) at \( M_t \) spatial locations, \( t = 1, 2, ..., 10, k = 1, ..., K \).

2.1 Model 1

Mendes et al. (2006) fitted the following spatio-temporal model for the hake recruitment data. Let \( R = \{R(s, t), s \in D \subset \mathbb{R}^2, t = 0, 1, 2, ..., 10\} \) be a hidden binary process that takes the value 0 if the observation \( z(s, t) \) is a structural zero, or 1 if \( z(s, t) \) is an observation (greater or equal to zero) coming from a Poisson distribution. The likelihood of the model is simply the joint distribution of the observed values \( Z(s_i, t) \) and the binary hidden process \( R \):

\[
L = p(Z, R) = p(Z|R)p(R).
\]

- **Level 1: Likelihood**

\[
z(s_i, t)|\lambda(s_i, t), R(s_i, t) \sim \begin{cases} 
\text{Poisson}(\lambda(s_i, t)), & R(s_i, t) = 1 \\
0, & R(s_i, t) = 0 
\end{cases} \quad (2)
\]

\[
R(s_i, t)|\theta(s_i, t) \sim \text{Bernoulli}(\theta(s_i, t)), \\
\lambda(s_i, t)|\alpha, \omega(s_i, t) \sim \text{Ga} \left( \frac{1}{\alpha}, \frac{1}{\alpha \omega(s_i, t)} \right),
\]

- **Level 2: Link Functions**

Logarithm and the logit of the processes \( \omega(s_i, t) \) and \( \theta(s_i, t) \) at location \( s_i \) and at time \( t \) are respectively linear functions of the covariates \( X_k(s_i, t), k = 1, ..., 5 \) and a hidden, unobserved Gaussian field. Specifically, let

\[
\gamma^{(\omega)} = (\gamma_1^{(\omega)}, ..., \gamma_5^{(\omega)})
\]
and
\[ \gamma^{(\theta)} = (\gamma^{(\theta)}_1, ..., \gamma^{(\theta)}_5) \]
be two sets of regression coefficients.

\[ \log(\omega(s_i, t)) = \gamma^{(\omega)} \mathbf{X}(s_i, t) + \eta_t(s_i), \]
and
\[ \logit(\theta(s_i, t)) = \gamma^{(\theta)} \mathbf{X}(s_i, t) + \eta_t(s_i), \]

where, \( \eta_t \) is an isotropic Gaussian random field, independent and identical in time. The priors for other parameters can be found in Mendes et al. (2006).

Inference on this model was carried using the WinBUGS software with relatively good predictions and model accuracy. Details can be found in Mendes et al. (2006).

2.2 Model 2

As an alternative model, we assume that the hake recruits \( Z(s) \) is a log Gaussian Cox process, hence we assume that conditional on \( \lambda(s) \), \( Z(s) \) are Poisson random variables with intensity function \( \lambda(s) \), where, \( \log \lambda(s) = \eta(s) \), \( \eta(s) = \mu(s) + \eta^*(s) \), \( \mu(s) \) is the mean field and \( \eta^*(s) \) is a isotropic Gaussian random field with mean 0 and covariance function

\[
\text{Cov} [\eta^*(s), \eta^*(s')] = \sigma^2 \eta \rho(|s - s'|) = \frac{1}{\tau^2} \exp(-\phi |s - s'|),
\]
for all \( s, s' \). However, we do not observe this process. What we observe at locations \( s_i \) is a thinned version with intensity given by

\[ \lambda_{\text{obs}}(s_i) = \theta(s_i) \exp[\mu(s_i) + \eta^*(s_i)], \]

where \( 0 < \theta(s_i) < 1 \) is the thinning probability, and \( \mu(s_i) = \mathbf{X}^T(s_i) \beta \).

Here, \( \mathbf{X}(s_i) \) are location specific covariances and \( \beta \) are regression coefficients.

Hence, we suggest the following hierarchical model.

- **Level 1: Data**
  
  \( Z(s_i) \) conditional on \( \lambda_{\text{obs}}(s_i) \) are independent Poisson random variables.

- **Level 2: Link functions**

  \[ \lambda_{\text{obs}}(s_i) = \theta(s_i) \lambda(s_i), \]
\( \theta(s_i) \) are independent and identically distributed random variables with beta \((a_1, a_2)\) density,

\[
\log \lambda(s_i) = \mu(s_i) + \eta^*(s_i),
\]

where

\[
\mu(s_i) = X^T(s_i)\beta.
\]

We further assume that at a location \( s_j \) where there is no observation, the mean of the gaussian field \( \eta(s_j) \) is given in terms of weighted sum

\[
\mu(s_j) = \sum_{i=1}^{M} k_{ij} \mu(s_i),
\]

where \( k \) is a spatial convolution kernel depending on some measure of distance between the \( j \)-th latent location and the \( i \)-th observation location. Typically \( k = \{k_{ij}\} \) is a \( N \times M \) non-stochastic matrix with elements \( k_{ij} \). We assume that the elements \( k_{ij} \) are given by the Gaussian kernel

\[
k_{ij} = \frac{1}{2\pi\xi^2} \exp \left\{ -\frac{d_{ij}^2}{2\xi^2} \right\}
\]

where \( d_{ij} \) represents the Euclidean distance between the \( i \)-th observation location of the study region and the location \( s_j \). We further assume that thinning is made independent of the intensity of the process.

- **level 3: Prior information on hyper parameters**

\( \tau^2, \phi, \beta, a_1, a_2, \xi \)

Inference on this model was carried using the \texttt{WinBUGS} software.

Note that the log Gaussian Cox process \( Z(s) \) is a non-stationary process with intensity

\[
\lambda^*(s) = \exp[\mu(s) + 1/2\text{Var}(\eta(s))]
\]

where \( \mu(s), \text{Var}(\eta(s)) \) are respectively the posterior mean and variance of the hidden Gaussian random field \( \eta \). (See, for example, Moller and Waagepetersen, 2004)

Hence, the expected abundance of hake recruits in an area \( A \), can be calculated or, at least, approximated by

\[
E[Z(A)] = \int_A \lambda^*(s) ds.
\]

Here, we do not know the functional form of \( \lambda(s) \), but we can approximate the integral up to a desired level of accuracy over a fine grid of points.
Acknowledgments: This work is partially supported by POCTI/MAT/44082/2002

References


A constrained maximum likelihood estimation for relaxing the exclusion restriction in causal inference

Andrea Mercatanti\textsuperscript{1}

\textsuperscript{1} Bank of Italy, Payment System Department, Via Milano 60/G, 00184, Roma, Italy, andrea.mercatanti@bancaditalia.it. The views expressed are those of the author and do not necessarily reflect the views of the Bank of Italy.

1 Introduction

The exclusion restriction is crucial in the identification of treatment effects in various causal inference methods. Historically, the assumption appeared in the literature concerning the instrumental variables method which has a long tradition in econometrics, and that has been applied in the context of causal evaluation. In particular, Angrist et al. (1996) showed that, under a suitable set of assumptions including the exclusion restriction, the nonparametric method of instrumental variables can identify causal treatment effects for compliers, the individuals who would receive the treatment only if assigned to it. Under a general approach to causal inference, labeled the Rubin Causal Model, the exclusion restriction requires that the instrumental variable has not a direct causal effect on the outcome. In term of a linear regression model this is equivalent to impose the absence of a probabilistic link between the instrumental variable and the error term.

Subsequently, research in causal inference turned from the nonparametric instrumental variables method to parametric models; in particular by the contribute of Imbens and Rubin (1997a) whose introduced a suitable likelihood function for the analysis of randomized experiments with noncompliance. The connection is in the fact that the randomized experiment with imperfect compliance is a template that can be adopted for the identification and estimation of treatment causal effects also in nonexperimental situations. For the instrumental variables model, the template is that of a randomized experiment with imperfect compliance in the sense that the particular instrumental variable adopted should have the role of a random assignment for which the treatment does not necessarily comply. In this parametric context, Imbens and Rubin (1997a) introduced a weak version of the exclusion restriction requiring that the assignment to treatment has to be unrelated to potential outcomes but only for noncompliers, the individuals that would receive or would not receive the treatment regardless of
whether it is offered.
In spite of its importance, the exclusion restriction can be often unrealistic in practice; however relaxing the assumption is not straightforward since it is directly related to the identifiability of the parametric models. Indeed, without the exclusion restriction, the parametric models do not have unique maximum likelihood points, but rather regions of values at which the likelihood function is maximized (Imbens and Rubin, 1997a; Hirano et al., 2000). Given this identifiability problem, previous studies propose to relax the assumption by relying on priors distribution in a Bayesian framework (Hirano et al., 2000), or by introducing auxiliary informations from pre-treatment variables in a likelihood based context (Jo, 2002).
The current study explores a new option, where we fully relax the exclusion restriction without introducing extra information respect to the usual set of conditions adopted to identify causal effect in the instrumental variable framework (Angrist et al., 1996).

2 The Model
Supposing a binary treatment and a normally distributed outcome, then relaxing the exclusion restriction introduces two mixtures of normal distributions in the general parametric model proposed by Imbens and Rubin (1997a). The corresponding likelihood function is indeed:

\[
L(\theta) = \prod_{i \in \varsigma(D_i=1, Z_i=0)} \omega_{a0} \cdot N(y_i|\mu_{a0}, \sigma_{a0}) \times \prod_{i \in \varsigma(D_i=0, Z_i=1)} \omega_{n1} \cdot N(y_i|\mu_{n1}, \sigma_{n1}) \\
\times \prod_{i \in \varsigma(D_i=1, Z_i=1)} [\omega_{a1} \cdot N(y_i|\mu_{a1}, \sigma_{a1}) + \omega_{c1} \cdot N(y_i|\mu_{c1}, \sigma_{c1})] \\
\times \prod_{i \in \varsigma(D_i=0, Z_i=0)} [\omega_{n0} \cdot N(y_i|\mu_{n0}, \sigma_{n0}) + \omega_{c0} \cdot N(y_i|\mu_{c0}, \sigma_{c0})],
\]

where \(\varsigma(D_i=d, Z_i=z)\) is the group of the units assuming treatment \(d\) and assigned to the treatment \(z\), \(\omega_{tz}\) is the mixing probability, that is the probability of an individual being in the \(t\) group, \(t = a\) (always-takers), \(n\) (never-takers), \(c\) (compliers) if randomly assigned to \(z\); \(\mu_{tz}\) is the outcome mean for the units in the \(t\) group and assigned to \(z\); \(\sigma_{tz}\) is the standard error for the units in the \(t\) group and assigned to \(z\).
However the estimation of mixed normal distributions models implies analytical and computational difficulties due both to singularities of the likelihood function and to the presence of several local maximum points (McLachlan and Peel, 2000). Moreover, in our context the analysis is complicated respect to usual studies on univariate normal mixtures. This is principally due to the switching of mixture components label indicators that complicates the identification of causal effects.
3 A restricted ML procedure

In order to resolve the analytical complications due to the presence of mixtures in (1), we proposed a constrained maximization procedure. The relevant constraining subspace can be easily identified by exploiting the information supplied by the usual instrumental variables set of assumptions, then without introducing any extra information. This requires reformulating (1) in order to make the likelihood as a function of the conditional mixing probabilities $\omega_{t|dz}$; the task is not difficult if taking into account the relationship $\omega_{t|z} = \sum_{d=0,1} (\omega_{dz} \cdot \omega_{t|dz})$, where $\omega_{dz}$ is the joint probability to be assigned to $z$ and to take the treatment $d$. The new formulation for (1) make possible to restrict the likelihood analysis to a spherical neighborhood of the vector $\hat{\phi}_{t|dz}$, whose elements are the estimated conditional mixing probabilities obtained by transforming the vector of the estimated marginal mixing probabilities $\hat{\phi}_t$ proposed by Imbens and Rubin (1997b).

In order to evaluate the relative merits of the restricted maximization procedure, we draw 100 samples of size 10000 from an hypothetical population violating the exclusion restriction, Table1. The maximum point interior to a spherical neighborhood of $\hat{\phi}_{t|dz}$ and having raw equal to 0.01, has been identified for each of these samples by running the EM algorithm. Table 2 reports some operating characteristics for the repeated estimates of the Compliers Average Causal Effect (C.A.C.E): $\mu_{c1} - \mu_{c0}$. The results are compared to those obtained by the standard ML method under the weak exclusion restriction, that is by imposing only in (1): $\mu_{a1} = \mu_{a0}$, $\mu_{n1} = \mu_{n0}$, $\sigma_{a1} = \sigma_{a0}$, $\sigma_{n1} = \sigma_{n0}$. We observe the estimations of the C.A.C.E. based on the restricted ML procedure systematically present better performances compared to those calculated by the standard method.

Table 1, Hypothetical population distribution.

<table>
<thead>
<tr>
<th>$t$</th>
<th>$\omega_t$</th>
<th>($\mu_{t0}, \sigma_{t0}$)</th>
<th>($\mu_{t1}, \sigma_{t1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>0.4</td>
<td>(0, 1)</td>
<td>(1, 1.2)</td>
</tr>
<tr>
<td>$n$</td>
<td>0.25</td>
<td>(1, 1.15)</td>
<td>(2, 1)</td>
</tr>
<tr>
<td>$c$</td>
<td>0.35</td>
<td>(6, 0.85)</td>
<td>(7, 0.7)</td>
</tr>
<tr>
<td>$P(Z_i = 1) = 0.25$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2, Performances of ML estimators for replications from the population presented in Table1 in estimating the C.A.C.E.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Mean bias</th>
<th>Root MSE</th>
<th>95% Interval Coverage Rate</th>
<th>Mean width</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE restricted to $\hat{\phi}_{t</td>
<td>dz}$</td>
<td>$1.1 \times 10^{-4}$</td>
<td>0.096</td>
<td>0.940</td>
</tr>
<tr>
<td>MLE under the weak exclusion restriction</td>
<td>0.051</td>
<td>0.111</td>
<td>0.912</td>
<td>0.368</td>
</tr>
</tbody>
</table>
References


Imbens G.W., D.R. Rubin (1997b); Estimating outcome distributions for compliers in instrumental variable models; Review of Economic Studies, 64, 555-574.


McLachlan G.J., D. Peel (2000); Finite mixture models; John Wiley and Sons, Inc.
Analyzing the performance of alternative unemployment estimators in small areas

Militino, A.F.¹, Ugarte, M.D.¹, González-Ramajo, B. ¹, López-Sagaseta, M. ² and Goicoa, T.¹

¹ Departamento de Estadística e Investigación Operativa, Universidad Pública de Navarra, 31006 Pamplona, Spain, militino@unavarra.es
² Instituto de Estadística de Navarra, Spain

Abstract: Small area estimation is specifically designed for estimating in areas where the sample size is so small that precise estimates can not be provided with direct estimators. In this case, the use of some kind of auxiliary information improves the precision. In this work we compare the behavior of several design-based, model-assisted, and model-based estimators using different auxiliary information to estimate the unemployment population in the seven regions (subdivisions) of the province of Navarra, Spain. To make the comparison, we perform a Monte Carlo simulation study drawing samples from the 2001 Census with the same design used in the Spanish Labour Force Survey (SLFS).

Keywords: Official statistics; auxiliary information; administrative registers.

1 Introduction

The SLFS is a quarterly survey of households living at private addresses in Spain. Its purpose is to provide information on the Spanish labour market that can then be used to develop, manage, evaluate and report on labour market policies. It is conducted by the Spanish Statistical Institute (INE). The target population includes all persons aged 16 or more living in private households. Yet there are multiple aims achieved with this survey, the estimation of unemployment population is one of the most relevant. The survey follows a stratified two-stage cluster design and, for each province, a separate sample is designed. The primary sampling units (PSUs) are Census Sections (areas with a maximum of 500 households) that are grouping in strata according to the size of municipality. In Navarra, in the first stage 91 PSUs are selected with probability proportional to size according to the number of households. For each PSU selected, a simple random sampling is applied to draw 18 households, and then, it inquires the overall residents of the household aged 16 or more (about 3000 people). This sampling design produces self-weighting samples at stratum level. In this work we check by simulation the benefits of using different kinds of auxiliary information.
in alternative estimators. The scenario is the same as the one used in the SLFS survey, but we draw samples from the 2001 Census. The auxiliary information considered in this study consists of using one or more of the following predictor variables: (E) ages-sex groups, (S) municipality sizes, (N) educational level, (P) unemployment populations register and (D) claimant employment status.

2 Design-based estimators

The design-based estimators calculated for every region \( d (d = 1, \ldots, 7) \) are direct, post-stratified, and composite estimators. We defined them below. The \textit{direct} estimator takes the form

\[
\hat{y}_d^{\text{direct}} = \frac{\sum_{j=1}^{n_d} w_j y_j}{\sum_{j=1}^{n_d} w_j} N_d = \hat{y}_d N_d,
\]

where \( y_j \) takes the value 1 for unemployed people and 0 otherwise. The weights \( w_j = \frac{P_h}{p_h} \), \( j = 1, \ldots, n_d \), with \( P_h \) is the total population among people aged 16 years old or more in the \( h \)-stratum and \( p_h \) is the total number of sampled people aged 16 years old or more in the \( h \)-stratum or province. \( N_d \) is the total population size in region \( d \), and \( n_d \) is the corresponding sample size.

The \textit{post-stratified} estimator is given by

\[
\hat{y}_d^{\text{post}} = \sum_g \hat{y}_{dg} N_{dg}, \quad \text{where} \quad \hat{y}_{dg} = \frac{\sum_{j=1}^{n_{dg}} w_j y_j}{\sum_{j=1}^{n_{dg}} w_j}.
\]

The \textit{synthetic} estimator is written as

\[
\hat{y}_d^{\text{syn}} = \sum_g \hat{y}_g N_{dg}, \quad \text{where} \quad \hat{y}_g = \frac{\sum_{j=1}^{n_g} w_j y_j}{\sum_{j=1}^{n_g} w_j}.
\]

Finally, the \textit{composite} estimator takes the form

\[
\hat{y}_d^{\text{dep}} = \lambda_d \hat{y}_d^{\text{post}} + (1 - \lambda_d) \hat{y}_d^{\text{syn}}
\]

where \( 0 \leq \lambda_d \leq 1 \), and is calculated as

\[
\lambda_d = \begin{cases} 1 & \text{si } \hat{N}_d \geq \alpha N_d \\ \frac{\hat{N}_d}{\alpha N_d} & \text{otherwise} \end{cases}
\]

\( \hat{N}_d \) is the estimated total size in region \( d \), \( g \) indicates the group, which is a combination of the categories of the considered auxiliary variables, and \( (\alpha = 2/3, 1, 1.5, 2) \).
3 Model-assisted estimators

We study the performance of the Generalized Regression (GREG) Estimators (Särndal, 1992) based on several models: a linear model, a logit model and a logit mixed model. If we consider a linear model given by

$$p_{id} = x_{id}^t \beta + \epsilon_{id}$$  \hspace{1cm} (1)

where $p_{id}$ is the probability that the $i$ th person be unemployed, $x_{id} = (x_{id,1}, x_{id,2}, \ldots, x_{id,p})^t$ is the vector of the $p$ covariates and $\epsilon_{id} \sim N(0, \sigma^2)$ are the random variables. The GREG estimator of the total number of unemployed population is given by

$$\hat{Y}_{d}^{GREG} = N_d \left( X_d \hat{\beta} + \frac{1}{N_d} \sum_{i \in s_d} w_{id} \left( y_{id} - x_{id}^t \hat{\beta} \right) \right)$$

If we consider the following logit model

$$\logit(p_{id}) = \log \left( \frac{p_{id}}{1-p_{id}} \right) = x_{id}^t \beta,$$ \hspace{1cm} (2)

the corresponding GREG estimator is

$$\hat{Y}_{d}^{GREG} = \sum_{i=1}^{N_d} \frac{e^{x_{id}^t \hat{\beta}}}{1 + e^{x_{id}^t \hat{\beta}}} + \frac{N_d}{N_d} \sum_{i \in s_d} w_{id} \left( y_{id} - \frac{e^{x_{id}^t \hat{\beta}}}{1 + e^{x_{id}^t \hat{\beta}}} \right).$$

Finally, the GREG estimator assisted in a logit mixed model of the form

$$\log \left( \frac{p_{id}}{1-p_{id}} \right) = x_{id}^t \beta + u_d$$ \hspace{1cm} (3)

where $u_d$ is the area random effect, is given by

$$\hat{Y}_{d}^{GREGMixed} = \sum_{i=1}^{N_d} \frac{e^{x_{id}^t \beta + u_d}}{1 + e^{x_{id}^t \beta + u_d}} + \frac{N_d}{N_d} \sum_{i \in s_d} w_{id} \left( y_{id} - \frac{e^{x_{id}^t \beta + u_d}}{1 + e^{x_{id}^t \beta + u_d}} \right).$$

4 Model-based estimators

We study the behavior of several estimators based on linear, logit, and logit mixed models. The estimator based on a linear model of the total number of unemployed population is given by

$$\hat{P}_{d}^{linear} = X_d \hat{\beta}$$ \hspace{1cm} (4)
The estimator based on a logit model of the total number of unemployed population is the following

$$\hat{P}_d^{logit} = \sum_{i=1}^{N_d} \frac{e^{x_{id}\hat{\beta}}}{1 + e^{x_{id}\hat{\beta}}}$$  \hfill (5)

Finally, the estimator based on a logit mixed model is

$$\hat{P}_d^{Mixed} = \sum_{i=1}^{N_d} \frac{e^{x_{id}\hat{\beta} + \hat{u}_d}}{1 + e^{x_{id}\hat{\beta} + \hat{u}_d}}$$

5 Bias estimation and mean squared error

We consider several measures of precision: the absolute value of the relative bias and its mean

$$SRA_d(\hat{y}) = \frac{1}{K} \left| \sum_{k=1}^{K} \hat{y}_d(k) - Y_d \right| \times 100, \quad SRAM(\hat{y}) = \frac{1}{D} \sum_d SRA_d(\hat{y}),$$

and the square root of the relative mean squared error

$$RECMR_d(\hat{y}) = \left( \frac{1}{K} \sum_{k=1}^{K} \left( \frac{\hat{y}_d(k) - Y_d}{Y_d} \right)^2 \right)^{\frac{1}{2}} \times 100,$$

and its mean

$$RECMRM(\hat{y}) = \frac{1}{D} \sum_d EMCR_d(\hat{y}).$$

6 Illustration and Results

From the 2001 Census, 500 simulations have been done under the same scenario as it is used in the SLFS survey. The auxiliary information considered in this study takes into account, individual or simultaneously, the 5 predictor variables defined in Section 1. The auxiliary variable (E) has 6 categories combination of age ($16 - 24, 25 - 54, > 55$) and sex, the stratum (S) gives the size of the cities or villages. It takes 6 values: (1) for the capital of the province, (2) for cities between 20000 and 50000 inhabitants, (3) for cities between 10000 and 20000 inhabitants, (4) for villages between 5000 and 10000 inhabitants, (5) for villages between 2000 and 5000 inhabitants and (6) for villages with less than 2000 inhabitants. The educational level (N) has two categories (1) for illiterate, primary or secondary school and
TABLE 1. Absolute value of the relative bias ($SRA_d$) and its mean($SRA$), the square root of the relative mean squared error ($RECMRM_d$) and its mean ($RECMRM$) for the direct and composite 4 (EP) estimators.

<table>
<thead>
<tr>
<th>Area</th>
<th>Direct SRA</th>
<th>RECMR</th>
<th>Composite 4 (EP) SRA</th>
<th>RECMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.716</td>
<td>58.456</td>
<td>6.620</td>
<td>28.389</td>
</tr>
<tr>
<td>2</td>
<td>0.874</td>
<td>34.226</td>
<td>2.966</td>
<td>16.312</td>
</tr>
<tr>
<td>3</td>
<td>1.539</td>
<td>31.806</td>
<td>4.153</td>
<td>16.255</td>
</tr>
<tr>
<td>4</td>
<td>0.399</td>
<td>33.294</td>
<td>1.576</td>
<td>17.568</td>
</tr>
<tr>
<td>5</td>
<td>1.716</td>
<td>30.023</td>
<td>8.325</td>
<td>18.300</td>
</tr>
<tr>
<td>6</td>
<td>0.522</td>
<td>19.342</td>
<td>3.393</td>
<td>10.973</td>
</tr>
<tr>
<td>7</td>
<td>0.859</td>
<td>8.652</td>
<td>3.072</td>
<td>7.180</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean</th>
<th>SRAM</th>
<th>RECMRM</th>
<th>SRAM</th>
<th>RECMRM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.232</td>
<td>30.828</td>
<td>4.301</td>
<td>16.425</td>
</tr>
</tbody>
</table>

(2) for technical workers and professionals. The unemployment population register (P) has three categories (1) occupied or inactive, (2) unemployed and (3) others. The variable claimant of employment (D) takes the value 1 if he/she is an employment claimant in the employment office of Navarra and 0 otherwise. The performance of the design-based, the model-assisted and the model-based estimators defined in Sections 2, 3 and 4 have been evaluated through the 500 Monte-Carlo simulations over 8 possible combinations of auxiliary variables: (E) edad-sex, (ES) age-sex-stratum, (EN) age-sex-educational level, (EP) age-sex- unemployment population register, (ESD), age-sex-stratum-claimant, (ESP) age-sex-stratum-unemployment population register and (ESN) age-sex-stratum-educational level. We choose the best estimator of the unemployed people by small areas between those with both minimum mean absolute value of the relative bias ($SRA$) and minimum mean square root of the relative mean squared error ($RECMRM$).

Table 1 shows the absolute value of the relative bias ($SRA_d$) and its mean($SRA$), the square root of the relative mean squared error ($RECMRM_d$) and its mean ($RECMRM$) for the direct and composite 4 (EP) estimators, in the 7 regions of Navarra for the direct and composite with auxiliary variables (E) and (P). The direct estimator is clearly unbiased but with large mean squared error for the whole province and for the small-areas. Results for direct estimator are showed for comparison purposes. There is no optimum estimator minimizing simultaneously $SRA$ and $RECMR$, but the best trade off between both criteria has been reached by composite 4 (design-based estimator with $\alpha = 2$).
Acknowledgments: Research supported by the Ministry of Science and Education (project MTM2005-00511) and OTRI project (2005 07 019).

References


Bayesian modelling of finite-source queueing systems

J. Morales\(^1\), M. E. Castellanos\(^2\), A.M. Mayoral\(^1\), R. Fried\(^3\) and C. Armero\(^4\)

\(^1\) Departamento de Estadística y Matemática Aplicada, Universidad Miguel Hernández, Avenida de la Universidad, s/n, 03202 Elche (Alicante), Spain
\(^2\) Departamento de Informática, Estadística y Telemática, Universidad Rey Juan Carlos, C/ Tulipán, s/n, 28933 Móstoles (Madrid), Spain
\(^3\) Departamento de Estadística, Universidad Carlos III, C/. Madrid, 126, 28903 Getafe (Madrid), Spain
\(^4\) Departament d’Estadística i I.O., Universitat de València, C/ Doctor Moliner, 50, 46100 Burjassot (València), Spain

1 Introduction

Queueing systems are a powerful tool for modelling and analyzing real life problems which involve congestion. The basic elements of a queue are the customers and the servers. When a customer arrives at the system and finds a free server, he or she is attended immediately. Otherwise the customer has to wait in queue until a server becomes available. Simple queueing systems are defined by the arrival pattern, the service mechanism including the number of servers, the distribution of the service times, the capacity of the waiting room, the size of the customer population and the discipline, that is the order in which the users in the queue are selected for service.

Real queueing problems with uncertainty are generally studied from data. Statistics becomes essential for describing the phenomena of interest and for drawing conclusions about its behavior. Bayesian reasoning is specially appropriate for queuing systems because we can evaluate posterior and predictive distributions for system performance measures taking different sources of uncertainties into account. Moreover, the Bayesian approach provides an optimal decision framework for design purposes due to the simplicity of the treatment of costs and losses (Armero and Bayarri, 1999).

This work focuses on a particular queueing system with servers, customers and spares. We discuss some predictive performance criteria for the operational capacity and availability of the system, providing design tools to improve the dimension of the service and the warehouse for spares.
2 The queueing model

The general setting examined here is commonly referred to as machine interference models (Gross and Harris, 1998) in the literature. They have originally been motivated by job-shops type systems (Gordon and Newell, 1967), but they have also been applied to many other real world problems in areas like telecommunication traffic, manufacturing and computer systems, transportation and semiconductor manufacturing. In particular, we consider an $M/M/c//r+s$ queueing system with a finite population of $r+s$ machines (users), $s$ of them treated as spares. All machines are subject to failures and thus maintained by $c$ repair crews (servers). One of $s$ spare machines, if available, can be substituted for a broken-down machine when being maintained. We assume that the operational times and the maintenance times vary independently according to exponential distributions with parameters $\lambda$ and $\mu$, respectively.

In general, the steady-state solution of a queueing system (when it exists) is expressed in terms of the distribution of the number of customers in the system. In our model, this distribution corresponds to the number of non-operative machines and always exists: the population of machines is finite and, consequently, the size of the queue cannot increase indefinitely. Availability measures defined in terms of operational, service and queueing times provide substantial information about the level of readiness of the system. The probabilistic distribution of all these congestion measures depends on the operational and maintenance rates $\lambda$ and $\mu$, usually unknown, and of the number of maintenance crews and spares.

3 Bayesian modelling and design

A full Bayesian formulation of the $M/M/c//r+s$ queueing model with parameter uncertainty is given by two levels:

- **Data level.** The data consist of a fixed number of operational and repair times, regarded as independent realizations of exponential distributions with parameters $\lambda$ and $\mu$, respectively

- **Prior level.** We use the natural conjugate choice, that is, a product of two independent Gamma distributions for the prior distribution of $\lambda$ and $\mu$.

The posterior distribution, $p(\lambda, \mu \mid data)$, is obtained from combining the prior with the likelihood. It is a product of two independent Gamma distributions as well, with updated parameters values.

This information allows us to develop a predictive analysis of the congestion of the queue in the steady-state which avoids any explicit reliance on values for the unknown parameters. For instance, we can compute the predictive
posterior distribution of the number, $N_{no}$, of non-operative machines, averaging the distribution of $N_{no}$ given $\lambda$ and $\mu$ with respect to (w.r.t.) the posterior $p(\lambda, \mu \mid \text{data})$. Analytic evaluation of this predictive will usually not be feasible, but approximation by Monte Carlo integration is always possible.

Other performance measures of a queue in a steady state combine operational and non-operational times, respectively, $T_o$ and $T_{no}$, to provide information on the machines ready for use at any moment (Kang et al., 1998). We examine the availability, defined as the global proportion of operational time:

$$A = \frac{T_o}{T_o + T_{no}}.$$  

The posterior predictive distribution of $A$ can be obtained as the expectation of the parameter-dependent steady solution $P(A \geq a \mid \lambda, \mu)$ w.r.t. the posterior distribution of ($\lambda, \mu$). Again, stochastic simulation can be used to approximate this predictive distribution.

The results mentioned above are very useful for design purposes. The queuing system can be designed by choosing a suitable number of servers and spares such that a certain acceptable system performance is achieved, with guarantees in terms of the predicted operational capability and/or availability. To this end, we discuss criteria concerning operational capability in terms of the predictive expectation of the number of operative machines $E(N_{o}|\text{data}) \geq r$ or the predicted probability of having at least $r$ machines in operation, $P(N_{o} \geq r \mid \text{data})$. For a profound discussion of these criteria see Morales et al., 2005. Similarly, some demand $a$ on the availability can be expressed in mean (predicted) terms, $E(A|\text{data}) \geq a$, or assessed by assuring a high (predicted) probability of achieving a large availability $P(A \geq a|\text{data})$.

Finally, a practical illustration of these results in an aeronautical scenario involving an airline with planes, repair crews for maintenance and spare planes will be discussed in the presentation at the workshop.

**Acknowledgments:** This research is partially supported by grants MTM2004-03290, GV05/018, TSI204-06801-C04-01 and MTM2004-02934.

**References**


Bayesian factor analysis applied to microarray data

D. Morris¹, P.J. Brown² and J.E. Griffin³

¹ UMR INRA-Agrocampus Rennes ‘Ecobiologie et Qualité des Hydrosystèmes continentaux’, 65 Rue de Saint Brieuc, CS 84215 - 35042 Rennes Cedex, France, David.Morris@rennes.inra.fr
² IMSAS, University of Kent at Canterbury, UK.
³ Department of Statistics, University of Warwick, UK.

Abstract: This paper proposes a method for the analysis of two-group comparative microarray experiments, where we aim to identify genes that are differentially expressed between two sources of tissue samples. We take a Bayesian approach to this problem and explore the use of factor analysis to represent the data in a lower dimensional space. The general factor analysis model is outlined, showing how it can be applied to microarray data. Then a Bayesian hierarchical model, which has sufficient structure to overcome the rotational problem associated with FA, is developed. We use prior distributions, with hyper-parameters shared across all genes, that induce shrinkage on the loadings vectors. As these loadings will highlight genes that discriminate between the two groups, the shrinkage will make interesting genes more outstanding. We demonstrate that there is enough information in the data to also sample model hyper-parameters using MCMC. This approach is shown, in applications to two data sets, to identify groups of similar observations, to highlight possible outliers and also to identify interesting genes.

Keywords: Factor analysis; microarrays; Bayes; MCMC.

1 Introduction

Microarrays are a relatively new technology that can measure the amount of mRNA relating to each gene within a given tissue sample. This is done on a scale much larger than ever before, where it is now possible to get the expression levels for thousands of genes simultaneously. The aim of these experiments may be to classify groups of covarying genes or observations using clustering techniques, or to identify genes whose expression levels have altered over different types of tissue samples. It is this second aim, of detecting differential expression, which we discuss in this paper. More specifically we focus on experiments comparing two cell types, where the exercise is to discover which genes are responsible for differences between the two pre-defined groups within the data. Often, for simplicity, it is assumed that the distribution of expression levels for all genes are independent, which can lead to both classical and Bayesian
Bayesian factor analysis tests for differential expression. These approaches inevitably run into multiple testing problems, which are often addressed using routines that try to control false discovery rates (Ge et al. (2003), Storey (2003)). However, with several thousand genes on an array, this can lead to very conservative procedures for detecting differentially expressed genes. To address this problem other approaches jointly model the distribution of expression levels, either in a hierarchical approach (Brown et al. (2004), Morris (2005)) or, as we propose here, using a latent structure that explains the co-variability between genes.

In this paper we assume that the data are a realization of some unobserved latent factors, where the number of factors is substantially less than the number of genes. These factors are likely to be related to the biological status of the cells from which mRNA has been taken for analysis. Our aim here is to find, within this lower dimensional representation, a projection which separates the observations into known groups. From this we can then identify genes responsible for the discrimination. A common approach to this type of problem is factor analysis (FA).

The following section will discuss approaches to gene selection and motivate the idea of jointly modelling the data. Then we discuss the classical FA model, followed by the Bayesian approach. A hierarchical model will be developed which has sufficient structure to overcome the classic rotational problems associated with FA. As the posterior distribution on model parameters is not in closed form, MCMC simulation is used for inference.

2 Gene-by-Gene Approaches

If we assume that the distribution of expression levels are independent for each gene, then a straightforward way to determine which genes are differentially expressed would be to preform classical t-tests. Two sample t-tests can be used for independent samples or with paired data we could perform paired t-tests for each gene. Genes are then selected as differentially expressed if the p-values are below some threshold. Accounting for multiple testing issues can be achieved using a variety of procedures to control the false discovery rate (Ge et al. (2003), Storey (2003)).

The independence assumption is convenient as it leads to simple tests for differential expression. However, as it is more reasonable to assume that genes do not act independently, methods which do not rely on this assumption are advisable. It is possible to adjust for multiple testing, while at the same time not making assumptions about the distribution of the t-statistics, using permutation approaches such as in Ge, Dudoit and Speed (2003). The distribution of a test statistic can also be estimated using Empirical Bayes methods as in Efron (2004). These approaches exploit the high number of genes in a typical microarray and therefore the large amount of information about the joint distribution of expression levels. Decisions can
then be made at the individual gene level, which take into consideration information from all the data collectively. Sharing information between genes is also an idea which suits a hierarchical modelling approach, to which Bayesian methods fit naturally. Prior beliefs can model the distribution of individual gene expression levels as arising from a gene wide population of distributions. Inference made for each gene then takes into consideration information from all other genes. This approach is taken in Brown et al (2004), by assuming that random effects at the gene level arise from some gene wide population, which leads to model hyper-parameters that are shared across genes. This approach was also used in Morris (2005), where additionally the proportion of differentially expressed genes is estimated from the data, leading to an automatic adjustment for multiple testing.

Hierarchical modelling has benefits over independence approaches, as information from all the expression levels affects decisions made for each gene. The idea of jointly modelling the data is therefore recommended as a better means to selecting a group of genes believed to be differentially expressed, than procedures which assume independence. In this paper we take a multivariate approach to modelling the dependance between genes, using a Factor Analysis model. The expression levels are assumed to be related through some latent factors that explain the observed variability in all genes, where the number of factors is much less than the number of genes. In our approach to Factor Analysis we specify hierarchical prior distributions, which again borrows strength across genes.

3 Factor Analysis

A factor analysis model assumes that multivariate data, having been observed in \( g \) dimensions, has arisen from \( k \ll g \) latent factors. The lower dimensional factors explain all the variability between samples and, after they are accounted for, leave only the variance specific to each dimension in the data. This idea fits nicely with microarray data where we have few observations in possible many thousands of dimensions, with genes as dimensions. The aim is therefore to represent each observation in a more useful low dimensional space which we assume reflects its biological status. The unobserved latent factors are referred to as factor scores in the following work.

The main issues when fitting a FA model are parameter estimation, rotational invariance and the choice of \( k \) (the number of factors). Here we will discuss model choice and rotational invariance. For a wider review see, for example, Mardia, Kent and Bibby (1979).
4 Bayesian Factor Analysis

With a Bayesian approach to factor analysis, the incorporation of prior information into the model naturally accounts for many of the problems associated with the classical approach. Our model is not just defined up to an arbitrary rotation. Instead, given informative prior distributions, we have a posterior probability distribution on the loadings matrix. The choice of $k$ also has a natural Bayesian solution. If we treat this as just another parameter then, with a suitable prior distribution, we will have a posterior belief in the number of factors. We can then make decisions based on the probability of $k$ given the data.

4.1 Prior Distributions

Previous approaches to Bayesian factor analysis can be seen, for example, in Press and Shigemasu (1989), Polasek (1997) and Lopes and West (2004). These approaches rely on a constrained loadings matrix to make the model identifiable. Here instead we suggest exploiting the amount of information in microarray experiments, to give us a well structured prior distribution. As there are typically many thousands of genes, we propose a hierarchical model where hyper-parameters are shared across genes. Shrinkage priors are used on the loadings matrix and mean vector as, a-priori, we may believe that only a few genes are responsible for changes in cell behavior. This approach can give useful results without having to put additional constraints on the model.

A normal-gamma prior (where the gamma distribution is on the precision, and hyper-parameters are shared across all genes) is placed on the loadings for each factor, which induces shrinkage in many of the elements. This prior has, marginally, a scale $t$-distribution which will lead to many small values while still allowing ‘interesting’ genes to show as large loadings. We may also have a normal-gamma prior on the elements of the population mean vector, depending on the data and our knowledge of the intercept term. If we have no knowledge about the mean then a vague prior may be appropriate, or alternatively we can estimate it by the sample mean. Prior distributions on the other model parameters also have a hierarchical structure.

4.2 MCMC

Our model does not have an explicit form for the joint posterior distribution of all the parameters. It is also not possible to calculate the marginal posterior for parameters of interest such as the factor scores, the loadings and the mean vector. However we can sample from the posterior distribution on all model parameters using Markov Chain Monte Carlo (MCMC) simulation. This technique has been used successfully in Polasek (1997), Lopes and
West (2004) and, with an application to microarray data, in West (2003). The MCMC sampler goes one step further here though than in these other papers. Rather than having to specify hyper-parameters lower down in the model hierarchy, we sample them within the algorithm. In our presentation, we will also discuss model choice, with reference to choosing the number of factors. A Simulation study is used to highlight some interesting features, such as multi-modality in the posterior, when fitting the incorrect number of factors.

5 Case studies

We have applied this approach to two data sets, one with independent samples and the other with paired data. For example here, we show results for two independent groups. In this case we have observations taken from two unrelated sets of tissue samples, with one of the groups acting as a control. Our aim is to discover which genes are differentially expressed between these two sources of data. The strategy will be to find a latent factor which discriminates well between the two groups. Genes are then identified as differentially expressed if they have relatively large loadings on this factor.

We fitted a four factor model to the data, where the second factor discriminated well between the two groups. The plots in Figure 1-(a) and -(b) show that the Bayesian FA model makes a clearer distinction between the two groups than a principal component (PC) analysis. The first eight observations have a genetic modification, while the last eight are the controls. We also present posterior means for the loadings on this second factor, and compare this with the eigen vector corresponding to the second PC. The Bayesian FA model makes interesting genes more outstanding.

6 Summary

This paper has developed a Bayesian approach to factor analysis, with the view to modelling latent structure within microarray data. The use of hierarchical prior distributions has given rotational information to the model and induced shrinkage onto the loadings matrix and vector of mean parameters. This model makes interesting genes more outstanding than empirical approaches such as principal component analysis. Sampling from the posterior distribution on all model parameters can been achieved through MCMC simulation. This has been shown to work well. The approach here has extended previous Bayesian approaches to factor analysis by not relying on a constrained loadings matrix and also making inference on hyper-parameters by sampling them within the MCMC algorithm. The model has been applied to two data sets, one with independent samples and the other with paired data. Results have shown that
the model is able to separate samples into known groups and identify genes responsible for the differences between the groups.

Acknowledgments: Most of this work was done while the first author was a student at the University of Kent at Canterbury, with sponsorship from the EPSRC and also the Sandwich base of Pfizer UK Ltd.

References


Bradley Efron (2004). Large-Scale Simultaneous Hypothesis Testing: The


Modelling the spatial dynamics of a population of trout using multistate capture-recapture methods

David Morris¹, Dominique Ombredane¹

¹ UMR INRA-Agrocampus Rennes ‘Ecobiologie et Qualité des Hydrosystèmes continentaux’, 65 Rue de Saint Brieuc, CS 84215 - 35042 Rennes Cedex. France, Davidlestermorris@hotmail.com, Dominique.Ombredane@agrocampus-rennes.fr

Abstract: In this study we apply a multistate (MS) capture-recapture model to a population of brown trout Salmo trutta, in a French coastal river system, to estimate survival and migration probabilities. We make inference on the different migratory tactics (within or out of the system) observed with juvenile fish along with survival rates, separating effects between different ages, seasons and geographic origins. With reference to the well known plasticity of the trout, we discuss how the use of covariates, such as the length of fish in their first year of life and population size, can help explain changes in movement. We also present results on the movement within each state (geographical section), given that a fish does not move between states, to add to the understanding of resident (fresh water) fish habitat usage. This original application of multistate capture-recapture modelling, to the dynamics of a fish population, will allow analysis of yearly changes in survival rates and of migration in terms of environmental impact and of human activities.

Keywords: Capture-recapture; multistate; ecological modelling; trout.

This study is part of a long-term investigation program on the ecology and dynamics of migratory salmonids in French coastal rivers. Recent work (Ombredane et al., 1998; Cucherousset et al., 2005) has shown that the brown trout (Salmo trutta) adopts a continuum of migratory tactics (from remaining in the brook where it is born to migrating to the sea). These tactics vary among years, caused by environmental variability and depend also on the growth rate, age and sex of the fish. The general goal of our particular research is to estimate survival and migration probabilities for different ages of the trout population of the river system Oir. In this paper we present some details of MS modelling and some first results from applying them to the trout data, where examples highlight issues with model fitting and diagnostic testing. These first results on data within a single brook allow us to make inference on the amount of
movement within the brook, given that a fish does not move downstream, and add to the understanding of resident fresh water trout habitat usage (Knouft and Spotila, 2002). We also discuss approaches to modelling the life history tactics of trout over the whole river system, which is the final aim of the project.

1 The trout data in the Oir river system

Eight cohorts of trout (about 7,500 individuals) from the Oir river system in Normandy (figure 1) were individually tagged, mostly at their young stage (6 months (0+ year) old), using Passive Integrated Transponders tag (PIT-tag). Including follow up recaptures, we have data from 1995 to 2003. As well as physical recapture (by electrofishing and in a trap localised in the downstream part of Oir), we also have data from stationary antennas installed in different points of the network to record continuously the passages of marked fish.

![FIGURE 1. The Oir river system, with areas of capture and recapture.](image)

The entire La Roche brook which is split up into 20 sections, of length varying from 35m to 167m, was electrofished in May and October each year. In the principal course of the Oir, only 5 sections (which represented more than 60% of the course) were prospected only in October from 1996. For the other brooks, only 2 short sections were electro-fished each October from 1996. So for each captured fish, at each time, it is possible to declare in which section it is captured and at what age and length. Therefore we have a life history for each trout tagged.

Data such as this fall under the scope of capture-recapture analysis and fit well into the popular, and developing, approach of multistate modelling. In our case, the states correspond to geographic sectors of capture and of recapture (Figure 1): La Roche brook, the principal course of the
Modelling the spatial dynamics of a population of trout in the Ori river (divided in 2 states) and a downhill trap leading to the Selune river/estuary/sea.

2 Multistate Capture-Recapture Models

Multistate (MS) models are an extension of standard capture recapture models (see Lebreton et al., 1992), with the additional feature that animals are allowed to move between predefined states during the study. The states are often taken to be geographical areas, as in this application, however they can also refer to physical attributes such as breeder/non-breeder status (Pradel and Lebreton et al., 1999). The models therefore include the same parameters as single state models, namely survival and capture probabilities. In MS models however we also have movement parameters, which define the probability that an animal stays within its original state or moves to another. The movement probabilities are conditional on survival and are defined by stochastic matrices, which leads to a matrix generalization of a product multinomial likelihood (Brownie et al., 1993).

Our approach to capture-recapture modelling follows the guidance in Lebreton et al. (1992), with further discussion on multistate approaches in Lebreton and Pradel (2002). We begin by testing the goodness of fit (GOF), which uses a series of Chi-squared tests implemented in U-care (Choquet et al., 2005), to detect deviances from the assumptions of MS capture recapture models. These include independence of individuals, lack of memory (Markovian) and an assumption that the fish we observe are representative of those we do not. The last point is not possible to test, however in our study capture rates are very high (>75%). After this diagnostic testing we begin the modelling process with the most general model, having separate survival, movement and capture probabilities for each time period and every cohort. From this we then try to reduce the number of parameters, in a top down process, in the search for a parsimonious model which can explain the main effects of interest. A combination of the AIC and likelihood ratio tests (LRT) are used for model choice and hypothesis testing. Reported parameters are therefore the maximum likelihood estimates from the final model selected.

The GOF tests specifically test departures from the assumptions of the JollyMove (JMV) model, which has capture probabilities that depend on the previous state the animal resided, rather than just the one the animal is actually captured in. In this case there is a memory effect on capture, which for our data is not appropriate. So our step-down model choice process begins with the conditional Arnason-Schwarz (CAS) model (Brownie et al., 1993), which has capture probability depending only on the current time and state. The CAS model is nested within the JMV and so the diagnostic test for JMV serves as the global test, in the absence of a specific test for CAS. For model fitting we have made use of the freely available software
M-Surge (Choquet et al., 2004), and for diagnostic checking we have used U-care (Choquet et al., 2005).

3 Modelling the movement of resident trout

As part of our larger scale study of the Oir river system in Normandy, we present here some results from the analysis of data from one of the supporting brooks (La Roche brook). The motivation for this came from the requirement to segregate the whole river system into clearly defined areas (the states) so that we can make inference on survival from each area and movement between the sites. We therefore must draw a line at certain places along the continuum, from the upper brook to the lower river, at points we consider define a change in habitat. This is achieved using ecological knowledge of the system together with statistical analysis of survival and movement within each area.

The 20 sections in the raw data are perhaps too many for a robust analysis, as all possible movements between sections could lead to a $20 \times 20$ matrix of parameters in any time period. We would prefer a more parsimonious approach and so begin by dividing the brook up into 4 states, each of roughly 500 m in length. We also discuss separating the Brook into 8-states, to detect movements over a much smaller range. However as the 8-state model includes many more parameters than the 4-state, we had to make more simplifications to make all parameters identifiable. This problem of balancing information in the data against the desire to answer ecological questions, highlights the importance of our aim to define well the different states along the river system in the larger study.

| TABLE 1. Sections of the La Roche brook. |
|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| State | A       | B       | C       | D       |
| Size  | 557m    | 586m    | 569m    | 479m    |
| Original sections | 1-6 | 7-10 | 11-15 | 16-20 |

The states, along with their respective lengths and a reference to the original sections they cover, is in table 1. The goal is to make inference on the amount of movement between these areas, and to determine whether this changes over time, age and season.

The choice regarding the number of sections to make up a single state was based on both the amount of information in the data, with respect to identifying all parameters, and previous ecological studies (Knouft and Spotila, 2002) which indicate that only a small percentage ($\approx 5\%$) of trout move further than 800 meters from their initial position. Due to boundary effects we can only be sure that a movement over at least two states signifies a
move over a minimal distance. Moving from state-B to state-C, for example, could be achieved by a fish at the boundary of these states who only moves a few meters. From B to D, however, signifies a move of at least 570m.

3.1 Diagnostic checking

The output from U-care gave an overall goodness of fit statistic of 88.33 on 101 d.o.f. (p=0.81), which shows no evidence to reject the general model. However on closer inspection of all tests included in this summary measure, we see that test 3G.SR fails at the 1% level. The null hypothesis for test 3G.SR can be stated as ‘there is no difference in the probability of being reencountered between newly marked individuals and those previously marked’. Here we make a two sided test, but a one sided explanation for failing this test is transience. If trout from outside the study area enter, are marked, then return to their original habitat, then this would make newly marked individuals have a lower probability of being seen again than previously marked fish.

A closer inspection of the specific Chi-Squared tests that make up the total in test 3G.SR, along with an examination of the data from the whole river system, did in fact lead to us discovering that a very small number of fish (8 out of 3539) were observed moving temporarily from the river to the brook over the winter, then returning in the summer. Therefore the fish marked for the first time in May have a probability of being transient.

The test 3G.SR is not specifically a test for transience, but the evidence presented does make us rethink our model with two suggested approaches. One parametrization has an age since first tagging effect on survival, so newly tagged fish are allowed to have different survival rates than previously tagged fish. Although this is a reasonable adjustment to the model, there is no GOF test available. In this case we must assume the model is suitable, given the data does not fail the other tests in U-Care. Our second approach removes from the data all fish tagged for the first time in May, for each of the years in the study. This will only slightly reduce the information in the data, as most initial markings are in October and we still have re-captures in May. For this reduced data set no problems were found with the GOF test (3G.SR now gives a p-value of 0.72).

As the general conclusions did not differ between the full data set with a 2-age model on survival, and the reduced data set with a standard model, below we only present results for the reduced data.

3.2 Model Fitting

The model fitting process proceeds by simplifying in areas where either we have prior knowledge of the system, or there is little information in the data. We do this to find a parsimonious model which can detect effects
of interest, and also to find evidence of consistency in parameters which may be useful for when we apply the model to the whole river system. Useful results for our main study were found here, including consistency in capture rates over the whole brook, however in this paper we only discuss the question with respect to seasonal and age effects on movement.

Table 2 displays summary measures (AIC, Deviance) for three possible models for the data, which differ in the way they model movement in Winter and Summer. The CAS model is also provided for reference. Models M2-M4 all have the same parametrization with respect to survival and capture probability, so the differences are purely attributable to movement parameters. We also have a two-age structure, where fish less than one year old have one set of survival and movement parameters, while all ages above one year share the same parameters. In an additional simplification for movement probabilities, we also share information across states. We assume that fish in the two middle states (B and C) behave roughly the same, as do fish in the two outer states (A and D). So the probability of moving to an adjacent state or moving two states is equal for states-B and -C. Also the probability of moving one, two or three states is equal for states-A and -D. So, for example, \( \psi_{AB} = \psi_{DC}, \psi_{BC} = \psi_{CB} \) and \( \psi_{BD} = \psi_{CA} \), where \( \psi_{ij} \) is the probability that a trout moves from state \( i \) to state \( j \).

<table>
<thead>
<tr>
<th>Mod</th>
<th>Descr’tn</th>
<th>AIC</th>
<th>P</th>
<th>Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>CAS</td>
<td>5240.6</td>
<td>191</td>
<td>4858.6</td>
</tr>
<tr>
<td>M2</td>
<td>age</td>
<td>5140.7</td>
<td>61</td>
<td>5018.7</td>
</tr>
<tr>
<td>M3</td>
<td>Season</td>
<td>5143.2</td>
<td>61</td>
<td>5021.2</td>
</tr>
<tr>
<td>M4</td>
<td>Age+season</td>
<td>5140.3</td>
<td>63</td>
<td>5014.3</td>
</tr>
</tbody>
</table>

**Table 2. List of models considered.**

The problem in identifying a season effect is that fish under one year old (0+ trout) are only observed in October (as they are born in spring), and importantly these are the majority of fish observed in this period (90% of the data). Older fish (1+ trout) have either died or emigrated from the study area. The number of fish that survive until they are 18 months old is relatively few compared to the new fish observed each winter. Therefore the only way to detect a seasonal effect is to compare 1+ movement between summer and winter. However this amounts to a relatively small part of the whole data set.

We can therefore fit three models which have similar weight, as judged by the AIC; an age only model with different movement parameters for 0+ and 1+ trout, a seasonal model (M3) where there are different movement parameters in October and May, and an age+season model (M4), which has movement parameters for 0+ trout (only in winter), movement for 1+ trout in Winter and 1+ in Summer.
The age+season model has the lowest AIC, although the difference is not large, and there is no evidence to reject the age model (M2) in favor of this on the basis of a LRT. The differences noted in this data are therefore explained more by age than season. So this fails to find evidence for a seasonal effect. We have also tried splitting the data further into 8 states, which does suggest a seasonal effect, with possible differences due to short distance movement (between 250-500m), which are too short to be detected with only 4 sections. While there were no GOF problems with the 8-state data, after again removing the May first time captures, there were many more problems concerning identifiability. The evidence for seasonal effects were based on perhaps over simplified models and we are currently examining some older data (captures from 1993 made only in the brook, and not for the whole river system) which may lend more support to this hypothesis.

In summary, we find that the probability for the 0+ trout to move over 570m is 0.05 (0.028,0.083) while for older fish it is 0.00 (0.00,0.015). Confidence intervals were achieved using the profile likelihood. Both movement estimates are for fish within the middle two sections, as we modelled the outer sections differently to take into account the marginal effect on leaving the system. This analysis gives useful quantitative results on resident trout movement using a model based approach, and includes differences due to age. Also the summary, that most resident trout spend their entire lives within 570m of their initial place of marking, is not in contrast to previous ecological studies.

4 Discussion and further work

The use of multistate modelling to describe the survival and movement of fish within river systems is still in its relative infancy. We have shown how this approach can be very useful for supporting ecological theories on the life history of fish, as well as obtaining useful estimates of changes in survival and movement probabilities over years. This is an original application of multistate capture recapture modelling, applied to the dynamics of a fish population, and shows potential to contribute to the understanding of trout survival and migratory tactics. Our continuing work applies these models to the life history of trout over the entire river system.

The life cycle of trout is well known to be plastic. Some fish migrate from the brooks, through the rivers, all the way to the sea. Other individuals however will not venture as far as the sea and spend their entire lives in fresh water. Ecological research has shown that the differences between types is not strongly related to genotype, therefore environmental conditions (Charles et al., 2005) have a strong effect on the life history of trout. As a contribution to the understanding of this area, we study the use of various covariates (such as population size, water flow rates and the length of trout caught...
in their first year of life) for first year movement and survival probabilities. Age is an important factor, as juvenile fish have very different survival and movement probabilities than adults.

In a two state model, for example, where the whole brook is now state-1 and the upper river (sections 4&5 in figure 1) is state-2, we have the probabilities of moving from the brook to the river over a four year period as \( \psi_{12}, \ldots, \psi_{42} \). With information on population size at time \( t \) indexed by \( x_t \), we have the simpler model \( \psi_{t12} = f(\beta x_t) \). The number of parameters required has reduced from four to one. In our study we have found evidence of this population density effect, with an analysis of deviance approach to variable selection giving a p-value of 0.009. This provides statistical backup for an ecological hypothesis, and also simplifies our model, hence we have more power to detect other features in the data. We also have the size of trout as a possibly covariate for our larger model, when estimating probabilities for migrating further downstream. So we can possibly use this information to simplify the model, perhaps together with the population density.

Acknowledgments: This research is funded by L’Institut national de la recherche agronomique (INRA), France.

References


Pradel, R. and Lebreton, J. D. (1999). Comparison of different approaches to study the local recruitment of breeders, *Bird Study (suppl.)*, 74-81.
Modelling threshold parameter as a function of covariates in segmented regression

Vito M. R. Muggeo

Dipartimento di Scienze Statistiche e Matematiche ‘S. Vianelli’, Università di Palermo, Italy, vmuggeo@dssm.unipa.it

Abstract: This paper is concerned with estimation of segmented regression models where the breakpoint of the piecewise relationship is not constant, but indeed depends on values of one or more variables. It is explored how a regression model for such parameter may be expressed and relevant estimation carried out. The method is illustrated on real dataset of daily time series of deaths, temperature and pollutant where the heat-tolerance is modeled as a function of particulate matter.

Keywords: Break-point; threshold modelling; segmented regression; interaction.

1 Introduction

Segmented regression is aimed to model nonlinear relationship between the response and a continuous variable $Z$, say, adjusting for additional possible confounders which have to be taken into account to avoid spurious relationships in the model. The nonlinear relationship of interest is approximated through a piecewise-linear curve, i.e. two or straight lines connected at one or more points within the range of $Z$ where the slopes change. Piece-wise modeling and relevant threshold find several applications in epidemiology, toxicology and risk assessment where the breakpoint is usually said ‘no effect concentration’ or more often ‘threshold’ (Schwartz et al., 1995; Pires et al. 2002).

While there are several papers in literature dealing with single breakpoint estimation, varying threshold is rather limited and not much used in practice, one possible reason being the existing difficulty in estimating the model: owing to the well-known non-regularity of such models, ad hoc algorithms have to be programmed and applied (Schwartz et al., 1995). In this paper we show how a regression model for the threshold may be set and estimation performed using standard maximization algorithms.

2 Methods

In simple segmented regression the piece-wise relationship between the response and the variable $Z$, is modelled via the term $(Z - \psi)_+$ in the linear
Threshold modelling

predictor; without losing in generality we can assume that the segmented variable $Z$ is defined in the range $[0, 1]$ and thus the threshold $\psi$ may be re-expressed via a logistic function:

$$\psi_i = \psi(\kappa) = \frac{e^\kappa}{1 + e^\kappa} \quad i = 1, 2, \ldots, n$$  \hspace{1cm} (1)

where $i$ is the index of the observations. When the breakpoint is unique (i.e. $\psi_i = \psi$ for any $i$), formula (1) is just a re-parameterization of the breakpoint $\psi$ in terms of $\kappa$, and it is substantially unnecessary in practice; however in a wider context, the above expression favours to be generalized to include the dependence on a covariate $X$ through the parameter $\kappa_1$:

$$\psi_i = \psi(\kappa; x_i) = \frac{e^{\kappa_0 + \kappa_1 x_i}}{1 + e^{\kappa_0 + \kappa_1 x_i}} \quad i = 1, 2, \ldots, n$$  \hspace{1cm} (2)

where $\kappa = (\kappa_0, \kappa_1)^T$. The main advantage in using a logistic function is that the parameters $\kappa$ are actually unbounded, making open the parameter space; in practice this implies that constrained optimization may be avoided. Model (2) says that the threshold value for the variable $Z$ is not constant but depends on the values of the variable $X$; the resulting threshold line is constituted by the pairs $(x_i, \psi_i)$ satisfying the (2).

For estimation we propose an iterative algorithm which generalizes the case of single breakpoint (Muggeo, 2003): the method relies on linearizing the nonlinear term $f(\psi(\kappa)) = (Z - \psi(\kappa))^+$, which is a composite function depending on two parameters via the logistic function (2). The procedure needs to evaluate at each iteration the first order Taylor expansion

$$f(\psi(\kappa)) \approx f(\psi(\kappa^{(0)})) + (\kappa - \kappa^{(0)})^T \frac{\partial f}{\partial \kappa} \bigg|_{\kappa = \kappa^{(0)}}$$  \hspace{1cm} (3)

where the $(j + 1)$th component of the gradient vector $\frac{\partial f}{\partial \kappa}$ is the $(j + 1)$th partial derivative, given by - using the chain rule:

$$\frac{\partial f}{\partial \kappa_j} = \frac{\partial f}{\partial \psi} \frac{\partial \psi}{\partial \kappa_j} = -I(Z > \psi) \frac{e^{\kappa_0 + \kappa_1 x_j}}{(1 + e^{\kappa_0 + \kappa_1 x_j})^2} x_j \quad j = 0, 1$$  \hspace{1cm} (4)

where $\psi$ is given by (2) and $x_0 \equiv 1$.

Details are skipped, but by algebra manipulation it may be shown that model having the threshold parameters according to (2), can be fitted by estimating iteratively a standard linear model including additional explanatory variables constructed by the expansion (3); such derived variables account for the estimation of parameter $\kappa$ whose point estimates and relevant approximate standard errors are available from the model output at the final convergence. Finally note that, unlike $\kappa_0$, it is reasonable to perform a test $H_0 : \kappa_1 = 0$ to assess whether the threshold location depends on the covariate $X$; at this aim the Wald statistic $\hat{\kappa}_1 / \text{SE}(\hat{\kappa}_1)$ can be employed.
3 Application

Temperature and air pollution represent serious threats for human health in urban areas; short-term effects of such environmental factors are well known and the following Poisson model is usually employed to analyze daily time series of deaths $Y_t$ as a function of seasonality (via a smoother $s(t)$), air pollution ($POLL_t$, linear term) and outdoor temperature $TEMP_t$:

$$\log E[Y_t] = s(t) + \delta POLL_t + \beta_1 TEMP_t + \beta_2 (TEMP_t - \psi)_+ \quad t = 1, 2, \ldots, n. \quad (5)$$

where the terms $TEMP_t + (TEMP_t - \psi)_+$ account for the V-shaped relationship between mortality and temperature and $\psi$ is the temperature value where mortality reaches its minimum; such value is understood to measure the heat tolerance since mortality increases as temperature goes beyond it.

In order to investigate on possible interaction of temperature-by-pollutant, we assess whether the threshold value of temperature can depend on pollution; that is, we replace the last term in the right hand side of equation (5) by:

$$\beta_2 (TEMP_t - \psi_t)_+ \quad \psi_t = (1 + \exp\{- (\kappa_0 + \kappa_1 POLL_t)\})^{-1} \quad t = 1, \ldots, n.$$

We apply the aforementioned method to daily time series in Palermo, South Italy, 1997-2001 using categorical variables month and years to control for long trend and particulate matter ($PM_{10}$) to quantify air pollution. Both pollutant and temperature have been evaluated as mean of lags 0-3. The

FIGURE 1. Estimated threshold line and 95% pointwise confidence intervals for the analysis of interaction effect of temperature and pollutant on mortality in Palermo.
fitted models with constant and varying threshold have Poisson deviance equal to 2148.4 and 2134.1 respectively; for the latter model, the estimated threshold line turns out to be:

$$\logit(\hat{\psi}_t) = 2.2889 - 0.0165 \text{PM10}_t$$ (6)

with standard errors equal to 0.2109 and 0.0034 respectively.

As discussed in the previous section, the Wald test $\frac{0.0165}{0.0034}$ provides evidence against a constant threshold yielding a very small $p$-value. Therefore results suggest that heat tolerance depends on pollution, namely high levels of airborne particles cramp heat tolerance, ranging from 28°C for low levels of PM10 up to 22°C during days characterized by high levels of pollution. Figure 1 displays the fitted threshold line with corresponding 95% pointwise confidence intervals.

4 Discussion

In this abstract it has been proposed a method to model dependence of threshold value as a function of continuous explanatory variables. To date, few paper have dealt with such topic in literature and practical applications are not frequent due to the important computational aspects. Here it has been illustrated a simple iterative algorithm which needs starting values only for the threshold parameters and can be implemented in standard statistical languages such as R, STATA or SAS. Simulations, here not shown, have emphasized that starting values are minor issues and the estimators are asymptotically unbiased with a limiting Gaussian distribution; also the relevant Wald test appear to guarantee the correct size.

Acknowledgments: The work was supported by PRIN Cofin MIUR 2004 prot. 2004137478_003.

References


Modelling genetic methods for data search strategies with ANOVA

Victor Muntés-Mulero¹, Marta Pérez-Cassany², Josep Aguilar-Saborit¹ and Josep-Lluís Larriba-Pey¹

¹ Departament d’Arquitectura de Computadors (DAC). Universitat Politècnica de Catalunya (UPC). C/Jordi Girona 1-3 08034 Barcelona, Spain, {vmuntes, jaguilar, larri}@ac.upc.edu
² Departament de Matemàtica Aplicada II. UPC, Spain, marta.perez@upc.edu

Abstract: Database management systems (DBMS) must find the most efficient strategies in order to retrieve the data required by user queries. However, with the increasing need for storing huge amounts of information, traditional optimization based on exhaustive search methods fail to find the optimal strategies. Genetic Programming has been proposed as an alternative random search strategy to improve the optimization quality in these situations. In this paper, we show that the ANOVA technique is useful to analyze DBMS optimization efficiency. We propose a statistical model fitted to 7290 observations that is able to predict the cost of the best result obtained by the optimization process, using an optimizer for DBMSs based on genetic approaches. Also, it offers the possibility to establish criteria to configure a genetic optimizer in order to obtain the best possible results.

Keywords: Query optimization; genetic optimization; ANOVA; nested factors; random effects.

1 Introduction

Large amounts of data pose special problems for Knowledge Discovery in Databases. Specifically, deciding the most efficient method to retrieve the data required by a user is still an unresolved problem, when the amount of relations involved in a single query is very large. In the Database Management Systems (DBMS) environment, this problem is usually referred as the large join query optimization problem [Ono, 90]. Special attention is payed to this issue by proposing new methods to find near-optimal search strategies for a fast data retrieval. One of the most recent proposals is based on genetic programming [Banzhaf, 98] search strategies. The suitability of evolutionary strategies has been shown to deeply depend on the problem to be solved. Due to the complexity of DBMSs, deciding whether an algorithm based on genetics is appropriate or not cannot be decided at a glance and demands a more formal statistical analysis.

393
In this paper, we show that the ANOVA technique [Scheff, 1959] [Montgomery, 1991] is very useful in this scenario. The linear model is based on 7290 observations, each one calculated from the average of 10 executions obtained under the same conditions, and it allows for 8 different factors. With this analysis, we prove that it is possible to find a model that predicts the average best cost obtained by a genetic optimizer. This implies that we can accurately study the effect of the different factors involved in the optimization process and set the base to establish criteria to parameterize a genetic optimizer.

In what follows, we describe the organization of the rest of the paper. Section 2 briefly describes our genetic optimizer. Section 3 presents the statistical model. Finally, Section 4 concludes the paper.

2 The genetic optimizer

Given a user query, typically written in SQL (Structured Query Language), a DBMS optimizer is in charge of deciding the best way to access the relations in the database, in order to retrieve the required information as fast as possible. The output of an optimizer is usually called query execution plan (QEP) (Figure 1). A query execution plan is commonly represented by a directed binary tree structure. The leaf nodes are the scan operations used to access the relations in the database, and the non-leaf nodes usually represent joining operations that merge data coming from two data sources. The root node returns the final results. In order to evaluate the goodness of a QEP, the optimizer uses a very complex cost function that express the fitness of the solution. The cost model used for this experiment can be found in [Muntés, Dec05].

![Query Execution Plan (QEP) example.](image-url)
As the number of relations involved in the query increases, the number of possible execution plan increases exponentially. Traditional optimizers cannot handle this situation since they are based on exhaustive search through the search space. Genetic optimizers, based on evolutionary strategies, have been proposed to remedy this situation.

The genetic optimizer used in this study is called CGO [Muntés, Apr05] and it is based on genetic programming. CGO uses the same execution patterns of any evolutionary algorithm. Starting with an initial population $P$ containing a known number of members denoted by $N$, usually created at random from scratch, two operations are used to produce new members in the population: crossover operations, which combine properties of the existing members in the population and mutation operations, which introduce new properties into the population. In order to keep the size of the population constant, a third operation, usually referred as selection, is used to discard the worst fitted members, using a fitness function. This process generates a new population, also called generation, that includes both the old and the new members that have survived to the selection operation. This is repeated iteratively through a fixed number of generations denoted by $G$. At the end, the best solution is chosen from the final population. We call $C$ the number of crossovers executed per generation and $M$ the number of mutations. Deciding the best value for each of these parameters is still an unresolved problem. The model presented in the next section offers a practical solution for deciding their impact and, therefore, establish criteria to parameterize a genetic optimizer.

3 General Statistical Model

In this section, we study the effect of different categorical variables on the average cost of the best QEP obtained after an execution by means of the Analysis of Variance technique. We have used the Statistical Analysis System (SAS) Release 8.00 [SAS, 99].

3.1 Variables in the model

Our statistical model aims at predicting the expected average cost of the returned QEP depending on different factors. As a consequence, the average cost of the returned QEP is the dependent variable. The factors used in the model are summarized in Table 1. Variables $N$, $G$, $C$ and $M$ have been fixed to take just three different values. In consequence, they can be considered factors. Our model must be general for any query and, therefore, it must consider some variables that define the main aspects of the query structure. Therefore, besides $N$, $G$, $C$ and $M$, we must consider two more independent variables of fixed effects which are: the number of Relations involved in the query ($R$) and the selectivity factor of the restrictions in the query ($S$). $S$
TABLE 1. *Independent factors in the model.*

<table>
<thead>
<tr>
<th>Factors</th>
<th>Description</th>
<th>Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td># of relations involved in the query</td>
<td>fixed</td>
</tr>
<tr>
<td>S</td>
<td>Selectivity of the query</td>
<td>fixed</td>
</tr>
<tr>
<td>Q</td>
<td>Query</td>
<td>random</td>
</tr>
<tr>
<td>N</td>
<td># of members in the population</td>
<td>fixed</td>
</tr>
<tr>
<td>P</td>
<td>Population</td>
<td>random</td>
</tr>
<tr>
<td>G</td>
<td># of generations</td>
<td>fixed</td>
</tr>
<tr>
<td>C</td>
<td># of crossovers</td>
<td>fixed</td>
</tr>
<tr>
<td>M</td>
<td># of mutations</td>
<td>fixed</td>
</tr>
</tbody>
</table>

is related to the probability of a tuple (or an item) in a base relation to qualify and to be returned as a result. Thus, queries with restrictions on the attributes that impose a selectivity factor close to 0 will discard, on average, most of the tuples in the base relation, and queries with selectivity factors close to 1 preserve most of the information, increasing the cost of the associated QEPs. It is reasonable to consider the level values for factor $N$ proportional to $R$, in consequence, $N$ is nested in $R$. Analogously, we consider the level values for $C$ and $M$ proportional to $N$ and, for that reason, $C$ and $M$ are both nested in $N$.

Due to the intrinsic randomness of the optimization process, it has sense to assume that there will be important differences in the cost for different queries with the same selectivity and the same number of relations. Thus, the query ($Q$) will appear in the model as a random effects factor. This allows us to consider the obvious effects of the query itself. The same applies to the initial population ($P$), and in consequence the model will contain two random effects factors.

3.2 Description of the experiments

All the decisions taken in this section are arrived at based on practical and empirical considerations from experiments done in previous ad-hoc tests, in

TABLE 2. *Independent fixed effects factor levels studied in the experiment.*

<table>
<thead>
<tr>
<th>Fixed effects</th>
<th>Independent Variables</th>
<th>Studied Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R$</td>
<td>20 and 50</td>
</tr>
<tr>
<td></td>
<td>$S$</td>
<td>$10^{-2}$, $10^{-4}$ and $10^{-8}$</td>
</tr>
<tr>
<td></td>
<td>$N$</td>
<td>$4R$, $8R$ and $12R$</td>
</tr>
<tr>
<td></td>
<td>$G$</td>
<td>50, 100, and 200</td>
</tr>
<tr>
<td></td>
<td>$C$</td>
<td>$\frac{N}{4}$, $\frac{N}{2}$ and $N$</td>
</tr>
<tr>
<td></td>
<td>$M$</td>
<td>$\frac{N}{4}$, $\frac{N}{2}$ and $N$</td>
</tr>
</tbody>
</table>
order to use reasonable and realistic values. Table 2 summarizes the levels used for the different factors in the experiment. Figure 2 depicts the experiment design. For every level in $R$ we randomly create 9 different queries ($Q$), three for each level of $S$. Given a query, we create 15 populations ($P$), 5 for each level of $N$. Once we have created the populations, we run 270 executions. This set of executions is divided into three subsets, corresponding to the three levels in $G$. For every level of $G$, we subdivide the executions into 9 subsets corresponding to the crossing between $C$ and $M$. Finally, for every possible configuration of the levels of each factor we run 10 executions and calculate the average cost obtained from them. Therefore, since we have run 72900 independent executions, the number of observations for the dependant variable is equal to 7290. The executions have been run on a distributed system with 80 nodes. Each node consist of an Intel Xeon 2.8 Ghz and 2 GB of RAM. The total amount of execution time used to obtain the results in this experiment is equivalent to more than 2700 hours in a single processor.

3.3 Model definition and Goodness of fit

We transform the average cost taking logarithm and we depart from the model that contains all the main effects and all the first-order interactions between the fixed effects factors. The first analysis of the model reveals the existence of 7 outliers. All the outliers correspond to configurations where the number of relations involved in the query is 50 ($R = 50$), the number of members in the population is the minimum tested ($N = 4R$), the number of generations is 50 ($G = 50$) and the number of crossovers and mutations is low. These results make sense since, under this configuration, we ask the genetic optimizer to obtain a good QEP in a huge search space, limiting its search potential by using the minimum number of members, generations
and operations tested and, therefore, it is possible that sporadically some observations stop far from a reasonably good QEP.
Once these outliers are removed, the data set has 7283 observations. The $R^2$ obtained with the same model and the new data set is 0.9693. After removing the interactions between $G$ and $M$, and between $C$ and $M$, both nested in $N$, which are not significative, the final model is (1). For this model the values of $R^2$, the Mean Square Error and the coefficient of variation are equal to 0.9691, 0.254 and 1.34, respectively. On the one side, this means that the model explains the 96.91 % of the total variability in the data. On the other side, the fact that the error standard deviation is quite small implies that the predictions will be accurate.

$$\log(Cost) = R + S + R \times S + Q(R \times S) + N(R) + S \times N(R) + G + P(Q \times N \times R \times S) + C(N \times R) + M(N \times R) + R \times G + S \times G + S \times C(N \times R) + S \times M(N \times R) + G \times N(R) + G \times C(N \times R)$$ (1)

Figure 3 shows us the relation between the observed and the fitted values. We can graphically see that our model is very good in predicting the average expected cost, depending on the different factors studied in this paper. Also, we can observe that the concentration of values near to 17 and 18 is higher than in other areas. This happens because the optimizer is trying to find a minimum cost. In our experiments, for queries involving 20 relations, the
minimum logarithm of the average cost seems to be around 17. The same effect can be observed for queries involving 50 relations, around 18. Figure 4 shows the studentized residuals versus the values fitted by the model. Out of the 7283 residual values, only 521 present an absolute value higher than 1.96, which represents the 7.15% of the total residual values. For the same reason explained above, in this plot we find two marked straight lines with a negative slope.

The hypothesis of independence of a linear model is guaranteed as a consequence of the fact that for a given population of a given query, we run independent executions. With respect to the normality assumption, studentized residuals have a larger kurtosis than the corresponding normal distribution with the same mean and variance.

4 Results and Conclusions

We defined a statistical model that allows us to study the importance of different factors in the optimization process of a DBMS. From the model, we extract the following conclusions:

- The ANOVA technique with fixed and random effects factors is useful to study and model the problem of optimization in DBMSs. We have
defined a model that is able to accurately predict the logarithm of the average cost obtained after the optimization process, using a genetic optimizer.

- Therefore, we can extract interesting conclusions that allow us to understand the keys to parameterize a query genetic optimizer from the model, which are independent from the random effects of the query and the population. Specifically, as an example, results show that all the factors considered in the model are statistically significant, meaning that they have a not negligible impact on the average cost. This observation is not trivial, since the contributions of the number of mutation operations executed per generation, for example, had not been studied before and, although it is quite clear that the use of this kind of operations is generally accepted, their impact for the query optimization problem had not been proved until now.

Acknowledgments: Special thanks to Calisto Zuzarte for his helpful comments and his support.

References


Modelling sample sizes of frequencies

Gerhard Neubauer¹ and Herwig Friedl²

¹ Institute of Applied Statistics, JOANNEUM RESEARCH, Graz, Austria,
² Institute of Statistics, Graz University of Technology, Graz, Austria

Abstract: Count data from registers are likely to exhibit underreporting and thus it is of high relevance to obtain an estimate of the true number of cases. One approach is to model counts as binomial variables with both parameters unknown. Estimation of the binomial denominator \( n \) is known to be problematic even in the simple iid case, if \( \bar{x}/s^2 \) is small. Several methods of stabilization, including a beta-binomial model, have been proposed to circumvent this. We suggest to use a regression model for \( n \) in the binomial and the beta-binomial case, and to estimate the parameters by maximum likelihood. Applying the proposed models to real data from the Austrian crime register we obtain estimates of the unknown total number of crimes, the so-called dark figure. The results of a simulation study show the good performance of the method.

Keywords: Binomial denominator; maximum likelihood; criminology.

1 Introduction

Since about two years there is a new register system for crimes in Austria called ‘Sicherheitsmonitor’ (SIMO) – ‘security monitor’ in English. It is an online system where every police station is connected to a central server, allowing for immediate electronic reporting of criminal activity (administration input side) and moreover for analytical queries on the whole nationwide database (analysis output). Besides the obvious advantage for everyday crime investigation, the data provide information for strategic planning in various aspects. One of these is regionalised prognosing, where on district level weekly counts of certain crime categories (larceny, burglary, robbery, assault, etc.) are predicted. Figure 1 shows an example of such data together with a smooth cubic spline fit. For the 120 Austrian districts there are many cases with a substantial change of criminal activity over time, resulting in a relevant trend curve like that in Figure 1.

For certain crime categories, where underreporting is a known problem, the estimation of the dark figure is of enormous interest. Dark figures can be estimated from survey data, but such surveys are expensive and current estimates are rarely at hand. A comparably quick and easy estimate becomes available when we assume a binomial distribution for a sample of such independent crime counts \( y_t \) observed in week \( t, t = 1, \ldots, T \). Let \( n \) be the number of crimes committed and \( p \) the probability of reporting it, then \( y_t \), the reported number of crimes follows a binomial distribution with parameters \( n \) and \( p \), both unknown. For data like those in Figure 1, constant
parameters are not appropriate. Thus, we allow $n$ to vary over time and obtain $y_t \sim \text{Binomial}(n_t, p)$. The dark figure is then $d_t = n_t - y_t$, motivating the need of any reasonable estimate of $n_t$. The interest in estimating the binomial denominator is not restricted to crime data only. Any register may be incomplete and in general any sample of count data is prone to estimation of $n$. Examples are unemployment data, infectious disease data, error counts in a production process, software engineering and the like. Usually $n$ is considered to be known and estimation concerns $p$ only. Approaches to estimate the unknown number of trials use the method of moments (MM) and maximum likelihood (ML) for either the binomial or a distribution derived under some mixing assumption for $p$. It is well known that under a binomial assumption both methods yield unstable estimates under certain conditions. For $y_t \sim \text{Binomial}(n, p)$, the MM estimates are $\hat{p}_{\text{MM}} = 1 - s^2/\bar{y}$ and $\hat{n}_{\text{MM}} = \bar{y}^2/(\bar{y} - s^2)$, where $\bar{y}$ and $s^2$ are the first two empirical moments. Both estimates are negative if $\bar{y} < s^2$, and their behavior becomes erratic if $\bar{y} \approx s^2$. The latter also holds for the ML estimates. Note that this instability does not affect the estimate of the mean $\hat{\mu} = \hat{n}\hat{p}$. Olkin et al. (1981) suggest to consider the quantity $q = \bar{y}/s^2$ to distinguish stable from unstable situations, where $q < 1 + 1/\sqrt{2}$ indicates instability. Besides that several methods to stabilize or robustify MM and ML estimation have been proposed (see e.g. Olkin et al., 1981; or Casella, 1986). Carroll & Lombard (1985) derive a ML estimate of $n$ from a marginal beta-binomial model for $y_t$, i.e. they consider $y_t|p \sim \text{Binomial}(n, p)$ and $p \sim \text{Beta}(a, b)$. Their rationale to this approach is to stabilize cases with small values of $p$. Theoretical investigations of the binomial likelihood show that the ML estimate of $n$ exists, but not its uniqueness (Olkin et al., 1981). Hall (1994), investigating the asymptotic properties of all three estimates, finds the Carroll-Lombard $\hat{n}_{\text{CL}}$ superior to the binomial MM and ML estimates. While the binomial estimates have heavy-tailed limits with infinite
moments, $\hat{\mu}_{CL}$ has light-tailed limits with finite moments.

Aitkin & Stasinopoulos (1989) propose canonical decomposition of the likelihood (Hinde & Aitkin, 1987) to eliminate the nuisance parameter $p$. A least-squares approximation of $L(n, p)$ gives separable functions $A(n)$ and $B(p)$, and inference can be based on $A(n)$ alone. Investigating a data example they find bad approximation using $L(n, p)$, but very good approximation using the reparameterization $L(n, \mu)$, where $\mu = np$.

Other work on $n$ estimation, especially Bayesian approaches, is not considered here (see Raftery, 1988; DasGupta & Rubin, 2004). Interesting, but limited to capture-recapture data, are the ideas in Van der Heijden et al. (2003), where a zero-truncated Poisson model is used.

2 Regression Models

A common feature of all previously mentioned work is, that only iid samples are considered. For independent but non-identically distributed data any of the above mentioned methods must fail, as the variability induced by such heterogeneity will lead to $\bar{y}/s^2 < 1$ in general. Thus, some regression approach is needed to model the denominators $n_t$ of such observations $y_t$.

Under a binomial assumption we have $y_t \sim \text{Binomial}(n_t, p)$. As positive values of $n$ are required, we consider the log-link model $\log(n_t) = x_t \beta$.

The MM approach is not appropriate for such a setting and thus ML is considered only. For the binomial model ML estimation is straightforward if we relax the restriction of integer valued $n$ and also allow real values. In order to keep the restriction $0 < p < 1$, we use its canonical representation $\logit(p) = \alpha$. Thus, the likelihood contribution of a single observation is

$$L(\alpha, \beta | y_t, x_t) = \left(\frac{\exp(x_t^T \beta)}{y_t}\right) p(\alpha)^{y_t} (1 - p(\alpha))^{\exp(x_t^T \beta) - y_t}. \tag{1}$$

Applying this regression approach to a beta-binomial model as proposed in Carroll & Lombard (1985) gives the likelihood contribution

$$L(a, b, \beta | y_t, x_t) = \left(\frac{\exp(x_t^T \beta)}{y_t}\right) \frac{B(y_t + a, \exp(x_t^T \beta) - y_t + b)}{B(a, b)}, \tag{2}$$

where $B(a, b)$ denotes the beta function with positive parameters $a, b$. Here, $p$ is considered a nuisance parameter and is therefore integrated out. However, information about $p$ is still available through $E(p) = a/(a + b)$ and var($p$) = $a/[(a + b)^2(a + b + 1)]$.

Carroll & Lombard propose to use some fixed values for $a$ and $b$ in their model, e.g. $(a, b) = (1, 1)$ or $(2, 2)$. For $(a, b) = (1, 1)$, the beta density shows a uniform shape, thus being an uninformative prior for $p$, whereas $(a, b) = (2, 2)$ puts more mass on the center. Moreover, $(a, b)$ can also be estimated by maximizing (2).
3 Austrian Crime Data

The following examples are based on data from SIMO. Weekly data are used on several crime categories in each of the 120 Austrian districts. We first label the data as of a binomial, Poisson, or negative binomial type by considering $q$ after adjusting the data for trend. The rule we used is: (i) binomial type, if $q > c$, (ii) Poisson type, if $1/c \leq q \leq c$, and (iii) negative binomial type, if $q < 1/c$, where $c = (1 + \sqrt{2})/\sqrt{2}$ (Olkin et al., 1981).

![Figure 2](image1.png)

**FIGURE 2.** Estimates $\hat{\mu}_t$ and $\hat{n}_t$ under a beta-binomial model for total crime.

![Figure 3](image2.png)

**FIGURE 3.** Estimates under the binomial model for the assault data.

After this classification we applied both the binomial and the beta-binomial model to the data. We further assume that the linear predictor $\sum_{j=0}^{4} \beta_j t^j$ sufficiently accounts for any trend structure over time. Estimation was implemented in SAS using PROC NLMIXED. The major advantage of this procedure is that it allows to specify any objective function to be maximized, like
TABLE 1. Estimation results for total crime and assault data

<table>
<thead>
<tr>
<th></th>
<th>Total crime data</th>
<th>Assault data</th>
<th>Assault data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>estim</td>
<td>s.e.</td>
<td>t</td>
</tr>
<tr>
<td>β₀</td>
<td>3.33</td>
<td>0.01</td>
<td>471.5</td>
</tr>
<tr>
<td>β₁</td>
<td>1.52</td>
<td>0.07</td>
<td>21.14</td>
</tr>
<tr>
<td>β₂</td>
<td>-14.27</td>
<td>0.37</td>
<td>-38.87</td>
</tr>
<tr>
<td>β₃</td>
<td>-10.75</td>
<td>0.65</td>
<td>-16.45</td>
</tr>
<tr>
<td>β₄</td>
<td>53.10</td>
<td>2.47</td>
<td>21.50</td>
</tr>
<tr>
<td>α</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>log a</td>
<td>1.28</td>
<td>0.17</td>
<td>7.65</td>
</tr>
<tr>
<td>log b</td>
<td>&lt; 0.01</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

the log likelihoods according to (1) or (2). Starting values for the iteration were set to β₀ = log(1.2 maxₓ(yₓ)), β₁ = · · · = β₄ = 0, and α = 0, or a = 1 while doing a grid search over b. In order to illustrate the behavior of both models when applied to data sets of different types, we concentrate on two examples. The first one is already displayed in Figure 1 and contains total crime frequencies in a touristic district. Because of q = 0.882, there is evidence that the data are of a Poisson type and that the binomial model is not appropriate. Hence, we maximize (2) resulting in Ŵ = 3.609 and õ = 1.001 giving an estimated mean reporting rate of Ũ(p) = 0.786. See also Table 1 for details on the estimates.

Figure 2 contains both, the estimated mean model ŵₜ = Ũ(p)Ŵₜ as well as ŵₜ. Similar to the non-parametric fit displayed in Figure 1, we can also see such a quadratic seasonal time trend in Figure 2, indicating touristical high season during summer time. The other example is on the number of assaults observed in a district surrounding a larger city. Here we observe q = 1.714 suggesting that the data are of a binomial type. Maximizing the binomial likelihood (1) gives the surprisingly small value ẑ = 0.379 (∆ = -0.494) for the estimated reporting rate of assaults. This may be still of a realistic magnitude when taking into consideration the known low reporting rate of domestic violence. Another interesting aspect of this data set is its rather robust behavior with respect to the maximization of the alternative beta-binomial likelihood (2). This again results in Ũ(p) = 0.379, but it is based on huge values of ẑ and ∆, indicating that the variance of the beta distribution assumed for the random quantity p is extremely small here. Besides that also the estimated means under both models are such that they cannot be distinguished in a graphical display. Table 1 contains the estimates under both models. Values for α indicate the block with estimates under the binomial model, while a and b are the corresponding estimates under a beta-binomial model. The results for β are so close under both models, that we state them only once in Table 1.
4 Simulation Study

To investigate the behavior of the proposed regression technique we did a simulation study based on the results from the analysis of the total crime data. We generated \( R = 500 \) samples each of size \( T = 50 \) using \( \logit(p) = 1.3 \) (i.e. \( p \approx 0.786 \)) for a binomial model, and \( \log a = 1.3, \log b = 0 \) (i.e. \( E(p) \approx 0.786 \)) for a beta-binomial model. For both cases we further assumed that \( \log n_t = \sum_{j=0}^{4} \beta_j t^j \) with \( \beta = (3.3, 1.5, -14, -11, 53) \). Both estimation methods were applied to both data situations. Tables 2 and 3 show the true parameter values, the Monte Carlo ‘mean’, and the respective estimates of standard errors. These are the Monte Carlo standard deviation of the \( R = 500 \) estimates, denoted by ‘sd’, and the Monte Carlo mean of the ML standard errors denoted by ‘se’.

Note that for the beta-binomial model \( \hat{p} \) is not uniquely defined by \( \hat{a}, \hat{b} \), as any pair \( a, b \) satisfying \( a/b = c \) gives the same \( \hat{p} = \hat{a}/(\hat{a} + \hat{b}) = c/(1 + c) \). \( \text{var}(p) \) allows to distinguish between solutions, and to assess the quality of estimates we consider both moments. There are two ways to estimate the reporting probability from the simulation results. First, as a function of the ‘mean’ estimates given in the tables, which we denoted by \( \hat{p}_1 \) or \( \hat{E}_1(p) \). For the binomial model \( \hat{p}_1 \) is just the logistic transform of \( \hat{a} \), whereas for the beta-binomial it is \( \hat{E}_1(p) = \hat{a}/(\hat{a} + \hat{b}) \) and \( \text{sd}_1(p) \). Second, as the Monte Carlo mean of the \( R \) replicates of \( \hat{p}_r \) denoted \( \hat{p}_2 \) and \( \hat{E}_2(p) \). Here we also get a variability estimate in the binomial case, denoted by \( \text{sd}(\hat{p}_2) \).

**Table 2. MC results for \( R = 500 \) simulated binomial samples**

<table>
<thead>
<tr>
<th>Setting ( \beta )</th>
<th>Binomial model</th>
<th>Beta-Binomial model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>sd</td>
</tr>
<tr>
<td>( \beta_0 )</td>
<td>3.3</td>
<td>0.086</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>1.5</td>
<td>0.155</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>-14</td>
<td>0.779</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>-11.0</td>
<td>0.913</td>
</tr>
<tr>
<td>( \beta_4 )</td>
<td>53.0</td>
<td>3.388</td>
</tr>
<tr>
<td>( \alpha ) 1.3</td>
<td>1.490</td>
<td>0.277</td>
</tr>
<tr>
<td>( \log a ) 1.3</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>( \log b ) 0.0</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

**Binomial Data** Maximizing the binomial likelihood given binomial data shows good results (see Table 2). The parameters in the predictor are close to their true values. Also both standard error estimates coincide quite good, indicating good performance of the method. The reporting probability estimates are \( \hat{p}_1 = \hat{p}_2 = 0.816 \) and \( \text{sd}(\hat{p}_2) = 0.203 \).

Utilizing the beta-binomial likelihood for these data gives quite similar parameter estimates. Although the estimated reporting probability
TABLE 3. MC results for $R = 500$ simulated beta-binomial samples

<table>
<thead>
<tr>
<th>Setting</th>
<th>Beta-Binomial model</th>
<th>Binomial model</th>
<th>mean</th>
<th>sd</th>
<th>se</th>
<th>mean</th>
<th>sd</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>3.3</td>
<td>3.270</td>
<td>0.024</td>
<td>0.011</td>
<td>5.908</td>
<td>3.477</td>
<td>4.851</td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>1.5</td>
<td>1.558</td>
<td>0.123</td>
<td>0.065</td>
<td>1.522</td>
<td>0.305</td>
<td>0.288</td>
<td></td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-14.0</td>
<td>-14.514</td>
<td>0.697</td>
<td>0.412</td>
<td>-14.208</td>
<td>1.533</td>
<td>1.496</td>
<td></td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>-11.0</td>
<td>-11.330</td>
<td>0.873</td>
<td>0.424</td>
<td>-11.030</td>
<td>1.837</td>
<td>1.645</td>
<td></td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>53.0</td>
<td>54.086</td>
<td>3.336</td>
<td>1.959</td>
<td>53.573</td>
<td>6.871</td>
<td>6.468</td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>1.3</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>-2.486</td>
<td>3.707</td>
<td>5.058</td>
<td></td>
</tr>
<tr>
<td>log $a$</td>
<td>1.3</td>
<td>1.625</td>
<td>0.215</td>
<td>0.168</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>log $b$</td>
<td>0.0</td>
<td>1.064</td>
<td>0.079</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
</tr>
</tbody>
</table>

$(\hat{E}_1(p) = 0.893$ and $\hat{E}_2(p) = 0.911)$ is larger than its true value, considering standard errors ($\hat{sd}_1(p) = 0.057$ and $\hat{sd}_2(p) = 0.180$) shows that this is not substantial. The estimates for $a$ and $b$ have large ‘sd’ values indicating insecurity about $p$, while ‘se’ compares to that of $\hat{\alpha}$ from the binomial model. Thus one should be cautious not to use the beta-binomial model for binomial data. Even though most things work well, there may result wrong inference.

**Beta-Binomial Data** Applying the correct model to the beta-binomial data gives reasonable estimates of the parameters in the $n_t$ model, as well as standard errors that are comparable to those for the correct binomial model in Table 2. Estimates of the reporting probability show some difference, because $\hat{E}_1(p) = 0.637$ whereas $\hat{E}_2(p) = 0.825$. But since $\hat{sd}_1(p) = 0.161$ or $\hat{sd}_2(p) = 0.171$ we conclude that this is not substantial.

For the binomial model there are still proper slope estimates, while the intercept is largely engrossed. This implies very large estimates for $n_t$. At the same time $\hat{\alpha}$ is negative leading to small reporting probability estimates: $\hat{p}_1 = 0.077$ and $\hat{p}_2 = 0.270$, with $\hat{sd}_1(p) = 0.444$. Also the standard errors for $\beta_0$ and $\hat{\alpha}$ are very large, indicating insecurity about the estimates. The binomial model gives parameter estimates such that $n_t$ and $p$ move towards their limits $\infty$ and 0, indicating the Poisson distribution.

5 Conclusions

The wide range of possible applications of $n$-estimation shows the high relevance of this topic. The proposed regression approach allows for greater flexibility than methods for iid data. Classifying data by $q$ into binomial, Poisson or negative binomial type of data helps to choose the appropriate method for $n$-estimation. In our examples the beta-binomial model turned out to be appropriate for data classified as Poisson type. The results of
the Monte Carlo simulations we did for binomial and beta-binomial data in the regression context are very promising and they clearly showed that the parameters in the system can be well estimated, provided that the appropriate model is used. However, there are still some situations where this approach fails. This is mainly the case when data are classified as negative binomial type and the binomial or beta-binomial likelihood is maximized. As our approach can be seen as an attempt to account for too much variability in the data, we hope to extend the regression models by additional elements capturing variability. An obvious candidate is to add some random effects in the linear predictor of the \( n_t \) model.

References


A comparison of SAS and R software for performing robust regression

Michael O’Kelly

1 Quintiles Ireland Ltd., East Point Business Park, Fairview, Dublin 3, Ireland, mokelly@quintiles.com

Keywords: Robust regression; SAS; R; documentation.

1 Introduction

Robust regression software offers a large number of choices, both of general approach and of parameterization within each approach. It is challenging to compare and indeed validate the offerings of the widely-used statistical packages for robust regression. This paper surveys offerings in R and in the new version of SAS, version 9. There are other options for robust regression software not described here. The rank-based method of Hettmansperger and McKean (1998) is available in Minitab. Koch et al. (1998) describe a nonparametric ANCOVA, robust to outliers in the response, using Cochran-Mantel-Haenszel test whose results can give an adjusted estimate of a treatment effect by inverting the test (see O’Kelly, 2003). Other software options are surveyed in Yaffee (2006).

It is desirable for robust regression methods to be efficient (have an acceptable rate of convergence to the true value of the parameters), and have a high breakdown value (where breakdown value measures the proportion of contaminated observations in a data set required to change the estimate of the parameters to any value from $-\infty$ to $+\infty$).

**Least trimmed squares (LTS)** estimation minimises the sum of the smallest $n$ residuals - relatively high breakdown but relatively inefficient.

**M-estimation** minimises a weighted objective function, which could be simply the usual least-squares function - relatively efficient, but relatively low breakdown point.

**MM-estimation** combines starting values for the regression coefficients from from LTS or from S-estimation (which has high breakdown, like LTS), an estimate of the scale from S-estimation, and performs the final estimation of the parameters via M-estimation - this combines the high breakdown properties of S-estimation, LTS and LMS with the efficiency of M-estimation.

Other robust regression methods available in SAS but not investigated further in this paper include least median squares (LMS) estimation (similar...
Robust regression: SAS and R

to LTS) and S-estimation, which minimises a function of the residuals, scaled by $s$, that gives the smallest $s$, has high breakdown, like LTS and LMS, but is somewhat more efficient. Although this is often used as part of MM estimation, no R package could be located that would give S-estimates. The three robust regression methods described above (LTS, M, MM) are applied to Brownlee’s (1960) stack loss data. Results from data on gross domestic product (Zaman et al., 2001) are also referred to for some comparisons.

2 Key results

2.1 Understanding and using the options available for the three robust methods

Different choices of options and tuning constants for the three methods can lead to quite different estimates of the location and scale of the response for a given data set. Among the choices available are

LTS: proportion of residuals to minimise; number of subsets to sample (for large data sets); number of ’concentration steps’ (see Rousseeuw et al., (2000)).

M-estimation: function of the residuals to minimise; tuning constant for that function; how to estimate the scale (may require another tuning constant).

MM-estimation: choice of a first, high-breakdown method of estimation (LTS or S-estimation, with their parameters); tuning constants for the final M-estimation.

The software documentation tends not to give a full account of how the robust regression is performed and how each of the tuning parameters influences the analysis. The user who wishes understand what the software is doing (and how to control it) must have access to the literature. As might be expected, SAS provides the fullest documentation, but still refers, for example, to Rousseeuw et al., (2000), Huber (1981) and Hampel et al. (1986) for some important items. While Venables and Ripley (2000) gives a concise introduction to the robust methods of their package MASS for the R language, much of the detail of the algorithms is left for the reader to figure out. The other R packages investigated - rrcov, roblm and robustbase - keep to the R tradition of minimal documentation, again assuming the reader will know the literature and thus understand the items that can be specified when running the package. Thus for example the roblm method allows the user to change tuning constants ”for the S-estimator” and ”for the re-descending M-estimator”, but does not give the equations into which these constants fit. This lack of full documentation is understandable, given the large number of steps in the complex algorithms used by robust regression, but can be frustrating for the user.
TABLE 1. Comparison of three robust regression methods using the stackloss

data.

<table>
<thead>
<tr>
<th>Robust method</th>
<th>Source</th>
<th>Intercept</th>
<th>Airflow</th>
<th>Water temp</th>
<th>Acid conc.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SAS: OLS</td>
<td>-39.9197</td>
<td>0.7156</td>
<td>1.2953</td>
<td>-0.1521</td>
</tr>
<tr>
<td>LTS*</td>
<td>SAS robust</td>
<td>-35.4078</td>
<td>0.8462</td>
<td>0.4453</td>
<td>-0.0924</td>
</tr>
<tr>
<td></td>
<td>rrcov</td>
<td>-35.4078</td>
<td>0.8462</td>
<td>0.4453</td>
<td>-0.0924</td>
</tr>
<tr>
<td></td>
<td>MASS</td>
<td>-35.2079</td>
<td>0.9129</td>
<td>0.2921</td>
<td>-0.1011</td>
</tr>
<tr>
<td>M-est.</td>
<td>SAS robust</td>
<td>-42.2853</td>
<td>0.9275</td>
<td>0.6507</td>
<td>-0.1123</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>8.2357</td>
<td>0.1177</td>
<td>0.318</td>
<td>0.1085</td>
</tr>
<tr>
<td></td>
<td>MASS</td>
<td>-42.2853</td>
<td>0.9275</td>
<td>0.6507</td>
<td>-0.1123</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>9.5316</td>
<td>0.1081</td>
<td>0.2949</td>
<td>0.1252</td>
</tr>
<tr>
<td>MM-est.</td>
<td>SAS robust</td>
<td>-41.710</td>
<td>0.9372</td>
<td>0.5743</td>
<td>-0.1129</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>7.8634</td>
<td>0.1126</td>
<td>0.2985</td>
<td>0.1039</td>
</tr>
<tr>
<td></td>
<td>MASS</td>
<td>-41.5230</td>
<td>0.9388</td>
<td>0.5795</td>
<td>-0.1129</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>9.307</td>
<td>0.1055</td>
<td>0.2879</td>
<td>0.1223</td>
</tr>
<tr>
<td></td>
<td>roblm</td>
<td>-41.5246</td>
<td>0.9389</td>
<td>0.5796</td>
<td>-0.1129</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>5.2978</td>
<td>0.1174</td>
<td>0.263</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>robustbase</td>
<td>-37.2631</td>
<td>0.8129</td>
<td>0.5344</td>
<td>-0.0712</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>3.0594</td>
<td>0.0644</td>
<td>0.1734</td>
<td>0.047</td>
</tr>
</tbody>
</table>

* (smallest 13 residuals out of 16)

2.2 Results from the SAS and R packages

For the stackloss data, Table 1 gives the results for the three robust regression
methods.
The R package rrcov and SAS PROC ROBUSTREG give identical estimates of the location parameters using LTS regression. The R package
MASS results differ, especially for the Water Temperature variable. The proportion of residuals chosen to be minimised had a considerable effect
on estimates of location. The author could not find a way to produce standard deviations of the location estimates directly from any of the packages
for LTS estimation. It is worth noting that, except for the MASS package, LTS software tends to emphasise what is referred to as the "weighted least
square" result. This result is simply an OLS estimate omitting outliers
and/or high-leverage points uncovered by the LTS procedure, but this fact is not always immediately clear from the packages’ outputs.

SAS and the MASS package had identical results for their M-estimates of location. The estimates of scale were close but not identical.

Finally, and perhaps surprisingly, given the large number of steps and options, there was good agreement for estimates of location between SAS, the MASS package and R’s roblm package, although they differed somewhat with regard to the estimate of scale. Results from the R package robustbase, however, did not agree with those from the other R packages or from SAS.

For Table 1, the final function minimised by the M-estimation was Tukey’s bisquare. For this function a tuning constant of $c=4.68501$ was used in the R packages and it was seen from roblm documentation that this corresponded to a choice of efficiency $\text{EFF}=0.95$ in SAS ROBUSTREG. For the first $S$-estimation a tuning constant of 1.54764 was chosen. This was the default in the R packages and was explicitly specified in SAS ROBUSTREG.

With the larger data set of Zaman et al., (61 observations, vs. 21 in the stack loss data set) the packages were in less close agreement with regard to both location and scale estimates for all robust methods. This may be due to the random sampling that is used for larger data sets by all methods.

3 Conclusions

For many of the approaches and options, the software packages give results that are fairly consistent. However, in some cases packages using procedures named identically according to the methods above give different results - sometimes quite different. None of the packages agreed with regard to estimates of scale. Because of this and because the algorithms are complicated, it is difficult to validate and feel confidently in control of current robust regression packages.

Acknowledgments: This research was sponsored by Quintiles Ireland Ltd.

References


Analysing the effect of breast cancer treatment on shoulder complexity: a multivariate random effects modelling approach

G. Oskrochi¹ and D. Shamley²

¹ School of Technology, Department of mathematical sciences, Oxford Brookes University, Oxford, UK
² School of Health and Social Care, Oxford Brookes University, Oxford, UK

Abstract: In this study muscle activity was detected by measuring the electrical current generated by muscles during an expansion/contraction. Four major muscles during scapulothoracic movement were investigated in patients treated with radiotherapy for breast cancer. The aim of the study was to relate the muscle activity to the patients’ report of pain and function.

Keywords: Multivariate models; random effects models; subject specific effect.

1 Introduction

Simple mastectomy and wide local excision (WLE) with irradiation of the remaining breast tissue are now standard procedures for the treatment of breast cancer. Despite the use of less extensive surgery there is still morbidity affecting the shoulder. The exact nature of the movement dysfunction is not yet known.

Researchers believe that radiotherapy has several known effects on parenchyma and vascular tissues. Connective tissue findings suggest that thickening of surrounding connective tissue may restrict movement of related tissues held within the confines of the band. Changes in the vascular network would cause muscle ischaemia whilst a limited ability to expand could have an effect on the efficacy of muscle contraction.

In this study muscle activity was detected by measuring the electrical current generated by muscles during an expansion/contraction. Four major muscles (i.e. pectoralis major, serratus anterior, latissimus dorsi and Rhomboid) during scapulothoracic movement were investigated in patients treated with radiotherapy for breast cancer. The aim of the study was to relate the muscle activity to the patients’ report of pain and function.
2 The Data

Patients who had been treated in the last 5 years for unilateral carcinoma of the breast without augmentation were included in the study. Prognostic variables (covariates) recorded included the treatment protocol they had received (simple mastectomy & irradiation; simple mastectomy, irradiation & axillary irradiation; WLE & irradiation; WLE, irradiation & axillary irradiation) and duration after surgery. Glenohumeral range was measured using the Fastrak motion analysis system and concurrent EMG readings were taken using Biopac for different angle of humeral elevation. Resting EMG was first taken for 2 mins and then patients performed 3 repeated movements of arm elevation to the point of pain (this was performed as scaption: $40^\circ$ anterior to the coronal plane). Each patient filled in the Shoulder Pain and Disability Index (SPADI) questionnaire. This is a series of visual analogue scales shown to be both valid and reliable for many shoulder conditions.

3 A Multivariate Repeated Measurements Model

We observe 4 measurements from each patient at each increment for about 200 patients. Hence a multivariate vector of observations at each increment. This can be regards as a multivariate repeated measurements model. $Y_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ y_{i4} \end{pmatrix} = \begin{pmatrix} y_{i11} & y_{i12} & \ldots & y_{i1m} \\ y_{i21} & y_{i22} & \ldots & y_{i2m} \\ y_{i31} & y_{i32} & \ldots & y_{i3m} \\ y_{i41} & y_{i42} & \ldots & y_{i4m} \end{pmatrix}$ for $i = 1, 2, \ldots, n$ The data shows strong association between $y_{ij}$ and $y_{ik}$, muscles of the same shoulder of the same individual. There is also strong association between $y_{ijr}$ and $y_{ij8}$, electrical activity of the same muscle in two different increments (repeated measures).

Let $f(y_{ij}|\tau_i)$ be a conditional density function of the $j^{th}$ muscle of the $i^{th}$ individual with unknown parameter vector $\beta_j$ given an unobserved individual specific (random) effect $\tau_i$, such that $\beta_j$ is related to the conditional linear predictor $\mu_{ij}|\tau = \beta_j X_i + \tau_i$, where $X$ is the observed covariate matrix and $\tau$ has distribution $p(\tau; \theta)$ with parameter vector $\theta$. Unconditionally, the marginal linear predictor has to be extracted from the unconditional distribution of the response i.e.

$$f(y_{ij}) = \int f(y_{ij}|\tau)P(\tau; \theta)d\tau$$

Conditional on the given value of the individual specific effect (random effects), the conditional distributions of $(y_{ij}|\tau_i)$ are assumed to be independent for $j = 1, 2, \ldots, 4$. 
Similarly, let \( f(y_{jk} | \upsilon_j) \) be the conditional density function of the \( k \)th measure of the \( j \)th muscles with the same unknown parameter vector \( \beta_j \) given an unobserved muscle specific (random) effect \( \upsilon_j \). Hence the conditional linear predictor \( \mu_{ijk} | \upsilon = \beta_j X_{ik} + \tau_i + \upsilon_j \). Similarly again, conditional on the given value of the muscle specific effect, the conditional distributions of \( (y_{jk}|\upsilon_j) \) are assumed to be independent for \( k = 1, 2, \ldots, m \).

Therefore, the contribution of the \( i \)th individual to the likelihood is

\[
L_i = f(Y_i) = \int \int \prod_j \prod_k f(y_{ijk}|\tau, \upsilon) p(\tau, \upsilon; \Sigma) d\tau dv. \tag{1}
\]

where \( p(\tau, \upsilon; \Sigma) \) is a bivariate random effect distribution with a general (unstructured) variance-covariance matrix. Univariate characterization of these issues (i.e. random effect or frailty models) has been discussed by many authors, including, Lancaster and Nickell (1980), Heckman and Singer (1984, 1985), Vaupel and Yashin (1985), Hougaard (1984, 1986a, b, 2000), Vaupel (1990), Aalen (1988, 1992), Richardson and Green (1997), and Verbeke and Molenberghs (2000). Many researchers in univariate case have chosen \( P(\tau; \theta) \) as the conjugate of \( f(Y; \beta|\tau) \) in order to get a tractable form for \( f(y; \beta) \). For instance, Clayton (1978), Lancaster and Nickell (1980), Clayton and Cuzick (1985), Crowder (1985) and Scallan (1987) based their models on gamma frailty; Whitmore and Lee (1991) based their model on inverse Gaussian distribution; Hougaard (1986a, b and 1987) based his models on positive stable distribution; and Richardson and Green (1997) based their model on Bayesian analysis of mixtures. Hougaard (1986a) extended the gamma distribution by an extra parameter to describe a family of strictly positive random variables. Aalen (1988, 1992) extended Hougaard’s (1986a) parameter range to include distributions with a non-susceptible subgroup (stayers). Oskrochi et al. (1997a) introduced a general compound distribution which covers most of the well known distributions. However, there is no guarantee that the use of conjugate distribution for unobserved random effects is the best choice. By using a non-conjugate distribution for frailty, the marginal distribution will analytically be intractable, therefore numerical integration or Monte Carlo simulation could be used. Alternatively, the integration may be approximated by analytically tractable forms. Researchers believe that it is important to place more emphasis on the correct specification of the model distribution than on the specification of the distribution of frailty. Therefore, one may simply use the central limit theorem to justify the use of a normal distribution as the distribution of the unobserved frailty when there is no prior knowledge about the nature of the frailty distribution. In a multivariate case, the normal distribution also allows for a general correlation structure between the random effects (Oskrochi and Davies, 1996). Maximisation of the likelihood in (1) is not an easy task and needs the specification of \( p(\tau, \upsilon; \Sigma) \). Although Van den Berg and Lindeboom (1995)
reported favourable experience with bivariate error distribution, however, the limited evidence to date on multivariate error distributions is less encouraging. In these circumstances, the nonparametric approach might not be the most appropriate for routine use with multivariate integrals. However, as highlighted earlier it is important to place more emphasis on the correct specification of the model distribution than on the specification of the distribution of the random effect. Hence, the use of multivariate normal distribution is by far the best choice when non-parametric frailty is not straightforward. This is also supported by the central limit theorem. In what follows we assume that \( p(\tau, \upsilon; \Sigma) \sim N(\Psi, \sum) \). Hence, without loss of generality we may also assume that the mean \( (\Psi) \) is absorbed in the systematic part of the model and \( p(\tau, \upsilon; \Sigma) \sim N(0, \sum) \). The multivariate numerical integration techniques such as multivariate Monte Carlo (MC) simulation, or the multivariate Monte Carlo Markov chain (MCMC) method could be used to bypass the analytical intractability problem. Fahrmeir and Tutz (1994), Oskrochi and Davies (1997), Oskrochi and Crouchley (2004) implemented the Cholesky decomposition for multivariate frailty models. The Cholesky decomposition decomposes the general variance-covariance matrix \( \sum \) into two triangular matrices, say, \( \Gamma \) and \( \Gamma' \) such that \( \sum = \Gamma \Gamma' \). This decomposition facilitates the evaluation of the multiple integral by Gaussian quadrature mass and weight. In general this approach assumes that the conditional distribution of the response, i.e. \( f(y_{ijk}|\tau, \upsilon) \), has a parametric closed form. For example, distribution from exponential family or GLM. In this particular case our transformed data supported a Normal distribution, i.e. we assumed \( (y_{ijk}|\tau, \upsilon) \sim N(\mu_{ijk}, \sigma_{ek}) \).

4 Empirical Results

Using a simple conventional model for parameter estimates and ignoring the exiting associations between the measurements leads to very unrealistic inference with almost all diagnostic factors (i.e. degree of the arm elevation, contraction or expansion move, duration since surgery, age, affected side, SPADI pain + SPADI disability, treatment types, ever received physiotherapy treatment, exercise within 6 months after surgery, and exercise now) become significantly important in explaining the electrical current generated by muscles during a contraction/expansion move. The following table shows the parameter estimates and the associated p-values when applying a conventional linear model to each muscle independently.
Not surprisingly the result is very misleading as the repeated observations are treated as independent observations. Analysing each muscle separately using a random effect model and ignoring the associations between muscles improves the parameter estimates substantially, but it is still not the most appropriate way of analysing this data because the muscles are assumed to be independent. The following table shows the parameter estimates of a random effects model applied independently to each muscle.

<table>
<thead>
<tr>
<th></th>
<th>PM</th>
<th>UT</th>
<th>SA</th>
<th>RM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree</td>
<td>.139</td>
<td>0.281</td>
<td>0.330</td>
<td>0.39</td>
</tr>
<tr>
<td>exp/cont</td>
<td>7.77</td>
<td>26.16</td>
<td>7.860</td>
<td>15.77</td>
</tr>
<tr>
<td>Dur aft surg</td>
<td>-.004</td>
<td>0.011</td>
<td>-.010</td>
<td>-.013</td>
</tr>
<tr>
<td>Age</td>
<td>.011</td>
<td>-.038</td>
<td>-.534</td>
<td>0.295</td>
</tr>
<tr>
<td>Affected side</td>
<td>-.00</td>
<td>-11.73</td>
<td>-1.97</td>
<td>-12.13</td>
</tr>
<tr>
<td>Spadi total</td>
<td>- .01</td>
<td>-.021</td>
<td>-.023</td>
<td>-.039</td>
</tr>
<tr>
<td>Mastecto</td>
<td>13.10</td>
<td>22.73</td>
<td>3.51</td>
<td>6.39</td>
</tr>
<tr>
<td>Martaxrt</td>
<td>13.40</td>
<td>15.84</td>
<td>23.30</td>
<td>11.98</td>
</tr>
<tr>
<td>Wlert</td>
<td>13.45</td>
<td>5.41</td>
<td>0.649</td>
<td>18.27</td>
</tr>
<tr>
<td>Wlaxrtrt</td>
<td>4.12</td>
<td>-8.44</td>
<td>-6.29</td>
<td>-7.28</td>
</tr>
<tr>
<td>wlxclrt</td>
<td>3.17</td>
<td>-5.26</td>
<td>-11.29</td>
<td>-2.08</td>
</tr>
<tr>
<td>Physio</td>
<td>-2.77</td>
<td>11.33</td>
<td>0.746</td>
<td>-1.58</td>
</tr>
<tr>
<td>Exercise1</td>
<td>-8.20</td>
<td>-22.11</td>
<td>-14.23</td>
<td>-11.61</td>
</tr>
<tr>
<td>Exercise 2</td>
<td>-8.73</td>
<td>9.50</td>
<td>-7.22</td>
<td>-16.40</td>
</tr>
<tr>
<td>Constant</td>
<td>7.88</td>
<td>18.80</td>
<td>62.50</td>
<td>14.73</td>
</tr>
<tr>
<td>$\sigma_v$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\sigma_e$</td>
<td>16.43</td>
<td>31.07</td>
<td>25.55</td>
<td>34.50</td>
</tr>
<tr>
<td>log Likelihood</td>
<td>-3910</td>
<td>-4500</td>
<td>-4319</td>
<td>-4597</td>
</tr>
</tbody>
</table>
Comparing the likelihoods of the two previous models confirms the presence of strong muscle specific effects ($\upsilon$). A test to assess the presence of individual specific random effect ($\tau$) returns a deviance difference of 480.79 for 1 degrees of freedom. Hence suggests a strong specific individual effect. Therefore a proper model is a model to allow for a joint distribution of the random effects with an unstructured variance-covariance matrix. Analysing the data with such multivariate repeated measures model leads to identification of real prognostic factors which are significantly affecting the muscle activity. The analysis considerably improves the inference based on simple random effects model, for example similar to the simple random effect model, the arm elevation and expansion/contraction are highly significant for the muscle activity. But, in contrast to the simple random effect model, affected side and total shoulder pain and disability index (SPADI) are both become significantly different from zero and both have shown adverse effect on the muscle activity. Inferences about these two factors are quite important for clinicians. On the other hand, a number of studies have looked at the effect of physiotherapy and exercise for the immediate and long term treatment of reduced shoulder function. Generally the results were that of no effect or a worsening effect. It is difficult to draw conclusions from those results as not all studies have given details of the exercise regime and once again measurement techniques and their analysis methods. Our analysis suggests that the physiotherapy and exercise does not have adverse effect on the muscle activity. A test to assess the homogeneity of the prognostic
factors on the four muscles i.e. $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4$ (excluding intercepts) will be rejected with deviance difference of 325.21 for 48 degrees of freedom. For instance the effect of the duration after surgery is quite different for pectoralis major and lattissimus dorsi compare to serratus anterior and Rhomboid. The effect of physiotherapy is not also the same for different muscles.

We are confident that this study and its analysis established the real cause of the shoulder dysfunction and therefore can lead to define a more suitable rehabilitation programme on these findings.

References


A nonlinear latent class model for joint analysis of multiple longitudinal outcomes and a clinical event

Cécile Proust¹ and Hélène Jacqmin-Gadda¹

¹ INSERM E0338, Université de Bordeaux II, 146 rue Léo Saignat, 33076 Bordeaux cedex, France

Abstract: In the context of cognitive ageing, we propose a latent class approach to explore the association between several correlated quantitative longitudinal markers and a clinical event. Thus, a nonlinear growth mixture model describes the latent classes of evolution of the global cognitive ability defined as the latent process underlying several quantitative psychometric tests, and the occurrence of a dementia is jointly predicted from the latent classes using a logistic model. By applying this model to data from the French cohort study PAQUID on cerebral ageing, we highlight the different profiles of cognitive decline associated with the occurrence of a dementia, and we provide diagnostic and prognostic tools based on the computation of the posterior probability of dementia using any available cognitive measures.

Keywords: Dementia; joint model; latent process; mixed effect model; mixture model; multivariate data

1 Context

In medical studies, both quantitative markers measured repeatedly and the onset of a chronic disease or a clinical event are usually collected. By joint modelling these two types of data, one can study the association between the markers evolution and the clinical event in order for example to describe the evolution of the markers during the course of a chronic disease (Jacqmin-Gadda et al., 2005) or to perform early detection of the disease using information on the markers evolution (Lin et al., 2002).

For studying such association between longitudinal markers and the clinical event, we propose a latent class approach which allows to investigate the heterogeneous structure of the markers evolution by highlighting latent profiles of evolution, and then to study the occurrence of the clinical event according to these latent profiles (Lin et al., 2000; Muthén and Shedden, 1999).

The latent class approach has several advantages over the shared random-effect approach which yet was more explored for studying association between longitudinal and survival data (Henderson et al., 2000). Firstly, it
relaxes the Gaussian assumption for the random-effects which is quite unrealistic when the population is heterogeneous. Secondly, interpretation of the relationship between the occurrence of the clinical event and the markers evolution is easier for clinicians since the latent class model exhibits profiles of evolution associated with the clinical event. Thirdly, in latent class models, the likelihood has a closed form and does not require any numerical integration.

Evolution of the disease is often measured by several correlated markers. For instance, in aging context, a battery of quantitative psychometric tests is collected to measure cognition. In order to make use of all the available markers when modelling the evolution of the disease, we propose to study the heterogeneous evolution of their common factor defined as a continuous-time latent process. Moreover, as markers can have different metrological properties (and often a distribution far from a Gaussian distribution), we use flexible parametrized transformations to link the markers with the latent process (Proust et al., 2006).

This work was motivated by the study of cognitive evolution in the elderly. By jointly modelling the heterogeneous evolution of the global cognitive ability, we can both describe the profiles of cognitive declines associated with dementia and propose an early diagnosis tool of dementia according to longitudinal cognitive measures.

2 Methodology

We consider $K$ correlated quantitative markers. For each marker $k$, $k = 1, \ldots, K$, each subject $i$, $i = 1, \ldots, N$, and each occasion $j$, $j = 1, \ldots, n_{ik}$, the measurement $y_{ijk}$ is collected at time $t_{ijk}$, $t_{ijk}$ being different for each subject and each outcome. The continuous-time latent process $\Lambda_i(t)$ represents the common factor of the $K$ markers. The link between the markers and the latent process is defined through a flexible monotone increasing transformation $h_k$ depending on a marker-specific vector of parameters $\eta_k$ to be estimated. For $h_k$, we chose Beta cumulative distribution functions which depend on two parameters and offer a large flexibility in the shapes (Proust et al., 2006). The measurement model links the measurement $y_{ijk}$ and the latent process at the time of measurement $t_{ijk}$ as follows:

$$h_k(y_{ijk}; \eta_k) = \Lambda(t_{ijk}) + \alpha_{ik} + \epsilon_{ijk}$$  \hspace{1cm} (1)

where $\epsilon_{ijk}$ are independent Gaussian errors and $\alpha_{ik}$ are random intercepts independently distributed according to a $N(0, \sigma^2_{\alpha_k})$ distribution.

The latent process evolution is then modelled through a linear mixed model including heterogeneous random-effects for investigating the unobserved latent class structure:

$$\Lambda_i(t) = Z_i(t)^T u_i + X_{1i}(t)^T \beta, \quad t \geq 0$$  \hspace{1cm} (2)
Nonlinear multivariate latent class model

\[ u_i \sim \sum_{g=1}^{G} \pi_{ig} N(\mu_g + X_{2i}^T \gamma_g, \omega_{2g}^2 B) \]  

(3)

where \( X_{1i}(t) \) is a vector of possibly time-dependent covariates associated with the vector of fixed effects \( \beta \). The vector \( Z_i(t)^T = (1, t, ..., t^p) \) is a time polynomial of degree \( p \) and the random-effect vector \( u_i \) is distributed according to a mixture of \( G \) Gaussian distributions with \( \pi_{ig} \) the probability that subject \( i \) belongs to class \( g \left( \sum_{g=1}^{G} \pi_{ig} = 1 \right) \).

By defining the \( G \) latent variables \( c_{ig} \) equalling one if subject \( i \) belongs to the latent class \( g \) and zero otherwise \( \left( \sum_{g=1}^{G} c_{ig} = 1 \right) \), we model the probability \( \pi_{ig} = P(c_{ig} = 1|X_{3i}) \) according to covariates using a multinomial logit regression (Muthén and Shedden, 1999).

For jointly analyse the occurrence of the clinical event, we modelled the probability of the event \( D_i \) (\( D_i \) equals 1 if the event occurs for \( i \) and 0 otherwise) by a latent-class-specific logistic regression as expressed by :

\[ P(D_i = 1|X_{4i}, c_{ig} = 1) = \frac{e^{\delta_{0g} + X_{4i}^T \delta_{1g}}}{1 + e^{\delta_{0g} + X_{4i}^T \delta_{1g}}} \]  

(4)

where \( X_{4i} \) is a vector of covariates and \( \delta_{1g} \), the corresponding vector of class-specific fixed effects. \( \delta_{0g} \) is the log-odds for the occurrence of \( D_i \) in the latent class \( g \).

A key assumption in this joint latent class model is the conditional independence of the markers evolution and the clinical event given the latent classes. Indeed, the latent class variable \( c_{ig} \) is considered as the only link between the longitudinal evolution \( (y_i) \) and the clinical event \( (D_i) \).

The likelihood of this joint latent class model has a closed form and is maximized using a Marquardt algorithm. Standard-errors of the estimates are obtained using the inverse of the Hessian matrix at the optimum.

3 Application

This work was motivated by the study of cognitive evolution in the elderly and its association with the occurrence of a dementia. The aim was to describe, during a follow-up of 12 years, the different profiles of evolution of the global cognitive ability, and to jointly predict the probability of dementia at the end of the follow-up according to these profiles of evolution. The global cognitive ability was defined as the common factor of three quantitative psychometric tests, the Isaacs Set Test, the Benton Visual Retention Test and the Digit Symbol Substitution Test of Wechsler.

From the Paquid study, a French prospective cohort study on cognitive ageing (Letenneur et al., 1994), we considered a sample of 834 subjects without dementia at the penultimate visit and with at least one measure per psychometric test during the follow-up ; the subjects had between 1
and 6 measures per test with a median of 3 or 4 measures per test. The common factor was assumed to have a quadratic evolution according to time until the 12-year follow-up and the model was adjusted on three covariates: gender, educational level and age. According to the Bayesian Information Criterion (BIC), two latent classes of evolution were highlighted. They are represented in Figure 1. The first trajectory had a very slight cognitive decline over time and an associated probability of dementia almost null (0.057) while the second trajectory was always under the first one with a sharper and nonlinear cognitive decline. In the latter, the associated probability of dementia was very high (0.938) showing that this latent class could represent the mean cognitive decline in a pre-diagnostic phase of dementia.

For early detecting a dementia according to the cognitive evolution, we proposed a diagnostic tool based on the probability of dementia according to the psychometric tests $y_i$ and the estimated vector of parameters $\hat{\theta}$:

$$P(D_i = 1|y_i; \hat{\theta}) = \sum_{g=1}^{G} P(D_i = 1|c_{ig} = 1; \hat{\theta})P(c_{ig} = 1|y_i; \hat{\theta})$$  \hspace{1cm} (5)

For comparing the performances of our multivariate model in detecting dementia with those of univariate models, we computed the area under the ROC curve (AUC) obtained from the estimated probability of dementia given in (5). In our sample, the AUC was 0.855 while it was between 0.775 and 0.808 when using only one psychometric test.
4 Concluding remarks

We proposed a nonlinear latent class model for modelling jointly three psychometric tests and diagnosis of dementia. In addition to the description of the two latent profiles of cognitive declines associated with the occurrence of dementia, we were able to propose a diagnostic tool and a prognostic tool of dementia. Especially, we highlighted that increasing information about cognitive evaluation by including several markers instead of a single increased the performances of the detection tool. Moreover, compared with simpler methods, this diagnostic tool had the main advantage that it could be used whatever the available longitudinal cognitive assessment on the three psychometric tests.

In this work, we focused on a multivariate approach for jointly model quantitative non Gaussian longitudinal markers and an event. For the moment we only considered the probability of occurrence of the event using a logistic regression which can imply a selection of the population studied. Further works will consist in extending this model to include a proportional hazard model for the time-to-event as proposed by Lin et al. (2002) when studying only one longitudinal marker.

References


Lin, H., Mc Culloch, C.E., Turnbull, B.W. et al. (2000). A latent class mixed model for analysing biomarker trajectories with irregularly scheduled observations. Statistics in Medicine, 19, 1303-1018


Appropriate statistical models for analysis of ratios

Pedro Puig

Departament de Matemàtiques, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain, ppuig@mat.uab.es

Abstract: In this talk we characterize all two-parameter continuous distributions that satisfy three properties that are suitable for analysis of ratios. An example of application is given.

Keywords: Characterization of distributions; generalized Inverse Gaussian distribution; harmonic distribution.

1 Properties of statistical models for ratios

Ratios of magnitudes are of considerable interest in Biosciences. Some of these ratios are, for instance, the Body Mass Index (Quetelet’s index), the Opsonic index, the ratio of systolic blood pressure at 3 minutes into recovery to systolic blood pressure at peak exercise, etc. More examples of ratios are certain index numbers in economics and the returns of stock market prices. The statistical models that will be considered here are sets of absolutely continuous positive random variables \( F_\theta \) whose probability density functions are indexed by the parameter vector \( \theta \).

We begin with some basic properties that, in our opinion, the statistical models describing ratios must satisfy:

1. Closure under change of scale. For any random variable \( X \) belonging to the model, \( \lambda X \) also belongs to the model, for any constant \( \lambda > 0 \).

2. Closure under reciprocals. For any random variable \( X \) belonging to the model, \( 1/X \) also belongs to the model.

Both properties can be reasonably assumed for most of the statistical models of ratios found in practice. For instance, suppose that we want to analyze a data set of currency exchange rates euro/dollar. For this example it is reasonable to assume that the scale is arbitrary (property 1), that is, the same statistical model that gives an accurate description of the currency exchange rates gives also an accurate description of the data set whose values are multiplied by 100. Moreover, it is also natural to expect that the same statistical model could describe the currency exchange rates dollar/euro, that is, the data set of the reciprocals (property 2).
Now we are also going to consider the following natural property concerning the estimation of the population mean:

3. The maximum likelihood estimator of the population mean is the sample mean.

From a practical point of view, it is desirable that the sample mean be the best estimator of the population mean. Scientists hold a very deep-rooted notion that the sample mean is the natural estimator of the population mean. Several authors have characterized distributions, basing their work on good properties that relate to the sample mean (Puig, 2003, Puig and Valero, 2006). For instance, it is known that the only location model (under mild conditions) such that the sample mean is the maximum likelihood estimator of the location parameter is the normal distribution. This is a long-established result from Gauss. Surprisingly, the class of two-parameter distributions satisfying these three properties can be determined.

2 Characterization

Theorem: Let $F_\theta$ be a two parameter statistical model with density functions $f(t; \theta)$ continuous in $\theta = (m, \Phi) \in R^2$ with support in $t > 0$. We also assume that these densities satisfy some regularity conditions. The model satisfies properties 1, 2 and 3 if and only if the density functions can be expressed as

$$f(t; \theta) = \frac{1}{2tK_0(\Phi)} \exp[-\Phi \left(\frac{t}{m} + \frac{m}{t}\right)], \quad t > 0,$$

where $K_0(.)$ denotes the modified Bessel function of the third kind with index 0.

The proof of this theorem can be found in Puig (2006). This family of distributions was first discovered by Halphen in 1941, who called it Harmonic law (see Seshadri, 1997). This model is also a particular subfamily of the Generalized Inverse Gaussian distribution (Jorgensen, 1982) and it has not very used in practice. However, this is a family with only two parameters very appropriate to analyze ratios.

3 An Example

We illustrate the use of the harmonic distribution on a set of stock market prices from the Spanish Stock Exchange records. They are the prices for Endesa covering the period from July 31 to November 30, 1998. They was analized in Puig and Stephens (2001) and the values are also given in

<table>
<thead>
<tr>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>3330</td>
</tr>
<tr>
<td>3285</td>
</tr>
<tr>
<td>3280</td>
</tr>
<tr>
<td>3215</td>
</tr>
<tr>
<td>3165</td>
</tr>
<tr>
<td>3215</td>
</tr>
<tr>
<td>3170</td>
</tr>
<tr>
<td>3095</td>
</tr>
<tr>
<td>3175</td>
</tr>
<tr>
<td>3190</td>
</tr>
<tr>
<td>3050</td>
</tr>
<tr>
<td>3065</td>
</tr>
<tr>
<td>3145</td>
</tr>
<tr>
<td>3290</td>
</tr>
<tr>
<td>3310</td>
</tr>
<tr>
<td>3035</td>
</tr>
<tr>
<td>3210</td>
</tr>
<tr>
<td>3210</td>
</tr>
<tr>
<td>3220</td>
</tr>
<tr>
<td>3180</td>
</tr>
<tr>
<td>3175</td>
</tr>
<tr>
<td>3370</td>
</tr>
<tr>
<td>3350</td>
</tr>
<tr>
<td>3345</td>
</tr>
<tr>
<td>3370</td>
</tr>
<tr>
<td>3395</td>
</tr>
<tr>
<td>3400</td>
</tr>
<tr>
<td>3445</td>
</tr>
<tr>
<td>3450</td>
</tr>
<tr>
<td>3555</td>
</tr>
<tr>
<td>3505</td>
</tr>
<tr>
<td>3410</td>
</tr>
<tr>
<td>3415</td>
</tr>
<tr>
<td>3445</td>
</tr>
<tr>
<td>3490</td>
</tr>
<tr>
<td>3500</td>
</tr>
<tr>
<td>3520</td>
</tr>
</tbody>
</table>

Table 1. Puig and Stephens (2001) considered the logarithms of the quotients of successive values (85 values), that is, the log-returns of the prices and they was analyzed by using the hyperbolic distribution (4 parameters) obtaining a good performance.

Here we analyze directly the quotients of successive prices. Usually the returns are assumed to be independent and identically distributed. This assumption is often a working hypothesis among statisticians in order to analyze returns or log-returns of stock market prices. The maximum likelihood estimator of the parameter of scale $\mu$ can be obtained directly from the expression $\hat{\mu} = \sqrt{\bar{x}/\bar{x}-1}$, where $\bar{x}$ is the sample mean and $\bar{x}-1 = \sum_{i=1}^{n}(1/x_i)/n$, that is, the sample mean of the reciprocals. The MLE of the parameter of shape $\Phi$ is obtained solving numerically the following equation:

$$ K_1(\Phi) = K_0(\Phi) = \sqrt{\bar{x}/(\bar{x}-1)}. $$

For the returns of Endesa in the considered period we obtain $\bar{x} = 1.00169$, $\bar{x}-1 = 0.99880$ and from here, $\hat{\mu} = 1.00145$ and $\hat{\Phi} = 2040.4$. The corresponding probability plot is shown in Figure 1. With only two parameters we have obtained a goodness of fit comparable with that of Puig and Stephens (2001) by using four parameters.

A $(1-\alpha)100\%$ approximate interval confidence of the population mean can be calculated from the expression, $\mu = \bar{x} \pm z_{\alpha/2}\sqrt{\hat{\sigma}^2/n}$, where

$$ \hat{\sigma}^2 = \hat{\mu}^2 + 2\hat{\mu}\hat{m}/\hat{\Phi} - \bar{x}^2. $$

For our data set, the 95% approximate confidence interval for the "trend" of the daily returns is $1.00169 \pm 0.00471$. From here, notice that there is no statistical evidence of an increasing trend for these returns.
Distributions of ratios

FIGURE 1. Probability Plot of the returns of ENDESA.

References


Stochastic simulation of the MAPK signalling pathway

Vilda Purutçuoglu¹ and Ernst Wit¹

¹ Lancaster University, Department of Statistics and Mathematics, Fylde College, Lancaster, LA1 4YF, U.K., v.purutcuoglu@lancaster.ac.uk, e.wit@lancaster.ac.uk

Abstract: The MAPK (mitogen-activated protein kinase) or its synonymous ERK (extracellular signal regulated kinase) pathway whose components are Ras, Raf, and MEK proteins with many biochemical links, is one of the major signalling systems involved in cellular growth control of eukaryotes including cell proliferation, transformation, differentiation, and apoptosis. Because of its diverse functionality, it is also activated in a variety of hormone activation and many illnesses, which have multi-complex gene structure like cancer and heart diseases. In this study we describe the MAPK/ERK pathway via (quasi) biochemical reactions and then implement the pathway by a stochastic Markov process. We use two techniques, Gillespie algorithm and diffusion approximation, to simulate reactions and we compare their results. A novelty of our approach is to use multiple parametrizations in order to deal with molecules for which localization in the cell is an intricate part of the dynamic process.

Keywords: MAPK pathway; stochastic simulation; multiple parametrization.

1 Introduction

The MAPK/ERK pathway whose main components are Ras, Raf, and MEK proteins (Figure 1), is one of the major signalling systems which regulates a variety of cellular activation from the cellular growth control of all eukaryotes to the activation of many fatal illnesses like cancer and heart diseases (Kolch et al. 2005). Therefore knowledge about the pathway is important both to understand the behavior of many biological activities and to produce the drug targetting true proteins of related illnesses. The structure of the MAPK/ERK pathway includes a number of phosphorylation on protein level whose interactions can be described by a stochastic process and directed by positive and negative feedback loops. Some of these phosphorylated proteins cause either activation or inhibition of other proteins. On the other hand, some of them, called scaffold proteins introduce high spatial dependency, affecting the signal flux by increasing the effective local concentration of components and enhancing their interactions (Kolch et al. 2005).

Current knowledge about the MAPK/ERK allows us to represent these biochemical activations as a list of (quasi) reactions. In this study we have described the pathway using 51 substrates and 94 reactions in which the 65
of them reflect changes in activities and translocation of substrates and the rest shows their degradations after dissociations. The following set of equations, which explains the activation of the pathway by the EGF receptor, is an example from the underlying list.

1. $\text{Grb2} + \text{SOS} \rightarrow \text{Grb2-SOS}$
2. $\text{EGF} + \text{Shc} \rightarrow \text{EGF} + \text{Shc}_m$ (translocation)
3. $\text{EGF} + \text{Grb2-SOS} \rightarrow \text{EGF} + \text{Grb2-SOS}_m$ (translocation)
4. $\text{Shc}_m + \text{Grb2-SOS}_m \rightarrow \text{Shc-Grb2-SOS}_m$
5. $\text{Shc-Grb2-SOS}_m \rightarrow \text{Shc}_m + \text{Grb2-SOS}_m$ (dissociation)
6. $\text{Shc-Grb2-SOS}_m + \text{Ras.GDP} \rightarrow \text{Shc-Grb2-SOS}_m + \text{Ras.GTP}$

in which Grb2, SOS, Shc, EGF, Ras.GTP, and Ras.GDP are single proteins, and Grb2-SOS and Shc-Grb2-SOS are protein complexes in the cytosol. In order to express the translocation of these substrates to the membrane, we have used the notation $m$. For instance the protein $\text{Shc}_m$ denotes the Shc protein translocated from the cytosol in to the membrane.

2 Stochastic modelling of the pathway

A stochastic process is a natural way to describe biochemical reactions such as the MAPK/ERK signalling pathway. However most of the currently available data are not suitable for this kind of dynamic modelling. Because of the lack of standardization, the available measurements depend on relative rather than absolute changes in concentrations without exact reaction rates (Kolch et al. 2005). Due to these limitations of the data, the common way to model gene regulations is to use the ordinary differential equations (ODE), which employ the law of mass action and the continuous concentration of each chemical species. Even though these models can
describe several reactions such as linear production and degradation successfully, they are unable to explain the stochastic nature of the actual interactions. Therefore we have implemented the MAPK/ERK pathway via a stochastic Markov process.

2.1 Gillespie algorithm

Under the assumptions that the rate law or hazard of molecular collision is constant in a small, fixed volume and in thermal equilibrium and also that the hazard of each reaction is proportional to the numbers of molecules of each reactant in a given reaction (the law of mass action), the biochemical system can be simulated by the Gillespie algorithm (Gillespie 1977). This exact algorithm generates consecutive pairs, consisting of the type and the time of the next reaction, from the exact probability of the associated master equations,

\[
\frac{\partial}{\partial t} P(Y; t) = \sum_{i=1}^{r} \left\{ h_i(Y - A_i, c_i) P(Y - A_i; t) - h_i(Y, c_i) P(Y; t) \right\}
\]

in which

\[
P(Y; t + \Delta t) = \sum_{i=1}^{r} h_i(Y - A_i, c_i) P(Y - A_i; t) \Delta t
\]

\[
+ \left\{ 1 - \sum_{i=1}^{r} h_i(Y, c_i) \Delta t \right\} P(Y; t).
\]

In these equalities, \( k \) stands for the number of molecule types \( Y_1, Y_2, \ldots, Y_k \) and \( r \) is the number of reactions \( R_1, R_2, \ldots, R_r \). The vector \( Y = (Y_1, Y_2, \ldots, Y_k) \) represents the state of the system at time \( t \), \( A_i \) denotes the \( i \)th row of the net effect matrix \( A \), and \( c_i \) is the stochastic rate constant of \( i \)th reaction, \( R_i \). Accordingly, \( h_i(Y, c_i) \) describes the hazard, and the term \( h_i(Y - A_i, c_i) P(Y - A_i; t) \Delta t \) describes the probability that the system is reaction \( R_i \) removed from state \( Y \) between \( t \) and \( t + \Delta t \) (Golightly and Wilkinson 2005).

Although this discrete event and continuous time algorithm works well for simulating small systems, it is inefficient for developing realistic complex models (Bower and Bolouri 2001, Golightly and Wilkinson 2005).

2.2 Diffusion approximation

By assuming that the probability distribution of the number of the molecules of each species at \( t \) depends on the continuous \( t \) and continuous number of molecules, the stochastic model is converted to the differential equations model. With the Fokker-Planck approach, the probability distribution
$P(Y(t))$ is expanded via a Taylor expansion. Then the change of state of each species at $t$ is found by Langevin approach, i.e. a diffusion approximation, in which a correlated noise term is added to maintain the stochastic behaviour of the model

$$dY(t) = \mu(Y, \Theta)dt + \beta^{1/2}(Y, \Theta)dW(t)$$

where $\mu(Y, \Theta) = A' h(Y, \Theta)$ and $\beta(Y, \Theta) = A' \text{diag}(h(Y, \Theta)) A$ are deterministic mean (drift) and variance (diffusion) matrices, respectively, depending on $Y$ and the parameter vector $\Theta = (c_1, c_2, \ldots, c_r)'$ explicitly. $dW(t)$ represents the change of a Brownian motion over time. As denoted in Gillespie algorithm, $A$ is the net effect matrix and $h(Y, \Theta)$ is the column vector of hazards $h_i(Y, c_i)$ (Bower and Bolouri 2001, Golightly and Wilkinson 2005).

## 3 Simulation

We have simulated the pathway by assuming that the hazards are constant for each level. We choose 3 gradations of reaction time speed, namely slow, normal, and fast. The stochastic rate constants have been calculated according to the order of each reaction, the given hazards and the number of molecules, which is initialized at 100 for all substrates.

We have run both algorithms under two different scenarios: excluding all degradation reactions and merely including EGF degradation, respectively. Indeed in biochemical reactions protein degradation is much slower than the time periods during which biochemical activation and de-activation processes take place. Therefore ignoring these reactions in the MAPK/ERK pathway is realistic in a simulation. Under the second scenario, we have added the “degradation” of EGF in order to simulate the effect of EGF dissociation from its receptor, which is a direct result of the activation of the MAPK pathway via the internalization into vesicles of this receptor.

We note that EGF is the only protein which triggers the activation of the pathway. In this way, we simulate the influence of this protein on the changes in activities of other substrates.

### 3.1 Updating plan of diffusion approximation

The algorithm starts by eliminating, initially, linearly dependent columns of the net effect matrix $A$. The positive definite diffusion matrix of the MAPK/ERK pathway is defined via 34 linear independent substrates. These 34 substrates are used to update the current state vector $Y_{\text{new}}$ which is the summation of previous $Y_{\text{old}}$ and the change $\Delta Y$ in a given small time interval, i.e. $Y_{\text{new}} = Y_{\text{old}} + \Delta Y$. 
3.2 Translocation and EGF degradation

Figure 2a and Figure 2b show the plots of total hazards of both Gillespie and diffusion through time under the two conditions. The diffusion seems a smooth approximation of the Gillespie algorithm. Furthermore in order to detect the effects of the translocation of proteins and compare the results of two conditions, we have plotted changes in activities of substrates with different localization versions through time. As observed in Figure 2c and Figure 2d, we have found that the localization of molecules by multiple parametrization is necessary for explaining the dynamic behavior of the system. Moreover we have seen that apart from several proteins in cytosol and near membrane like any types of single/complex MEK, ERK, c-Fos, and RKIP, most of the proteins are affected significantly by the change in EGF (Figure 2e and Figure 2f). These effects are varied. For instance, the number of molecules of active Ras (Ras.GTP), Raf (without making any complex with RKIP), Shc, Grb2, SOS, and of any kind of their complex structures decrease near membrane with the degradation of EGF, whereas they increase their number in cytosol.

3.3 Dynamic profiles

For observing homogeneous subsets between two algorithms, initially, we have clustered the proteins by principle component analysis. The results from both methods show that inactive Raf (Raf.I), active MEK (MEK.p2), c-Fos phosphorylated by ERK (c-Fos.p), and RKIP phosphorylated by either PKC or ERK explain most of the variability of the network. Indeed a similar conclusion is reached in the study of Hornberg et al. (2005) such that only a small group of reactions, which consist of Raf, MEK, and ERK, control the majority of pathway. We applied clustering (similarity matrix = correlation, number of cluster = 8) and obtain Figure 3. The results with EGF degradation indicate that both simulation techniques partition most of the proteins in the same clusters. However the diffusion approximation have more correlated proteins than Gillespie algorithm. The reason is that the former is much smoother than the latter.

References


FIGURE 2. Total hazards of Gillespie algorithm (black line) and diffusion approximation (grey line) by (a) excluding all degradation of EGF and (b) including only the degradation of EGF. Change in activities of Grb2-SOS complex (c) in cytosol and (d) near membrane, respectively including the degradation of EGF by using both Gillespie algorithm (black line) and diffusion approximation (grey line). Change in activities of (e) active Raf near membrane and (f) active ERK in cytosol, respectively including both with EGF degradation (black line) and without EGF degradation (grey line) by using the Gillespie algorithm.
FIGURE 3. PAMSAM clustering of 45 substrates (excluding c-Fos.DNA, GAP, PP2A, PAK, PP5, PKC) by using (a) Gillespie algorithm and (b) diffusion approximation when the degradation of EGF is included.


Identification of level shifts in stationary processes

William Rea¹, Marco Reale¹, Carmela Cappelli² and Jennifer Brown¹

¹ Mathematics and Statistics Department, University of Canterbury, Christchurch, New Zealand
² Dipartimento di Scienze Statistiche, Università di Napoli Federico II, Naples, Italy

Abstract: ART was introduced as a modified use of CART methodology for quick detection of structural breaks in the mean levels. In this paper simulations are presented to test ART against a number of different types of time series, to find a good pruning method, and to compare with alternative approaches.

Keywords: Time series; regression trees; structural breaks.

1 Introduction

Artheoretical Regression Trees (ART) (Cappelli and Reale, 2005) is a simple and fast approach to detect structural breaks in the mean in time series. It is an ordinary regression tree procedure where the dependent variable is the time series under consideration and the covariate is a strictly monotonic positive or negative sequence. In this paper a simulation study is presented to further investigate strengths and weaknesses of the method and compare it to the established procedure proposed by Bai and Perron (BPP) (1998, 2003)

2 Simulations

Simulations were run with series of uncorrelated observations drawn from an N(0,1) population with a single break at the mid-point giving two regimes. There where 16 regime sizes, 5² to 20² observations in length, with break sizes ranging from 0.05 to 2 standard deviations in steps of 0.05 standard deviations and 1,000 replications of each combination of regime length and break size. Our results indicate ART performs well, with α = 0, when the regime length is long and the break size is large.

Figure (1) shows the results of one simulation to test ART’s ability to correctly locate a single break in a series. These results show the break size is the important criteria.

438
This set of simulations exposed a problem with ART finding substantial numbers of spurious breaks when the regime length is small and using the default, $\alpha = 0$, cost-complexity pruning. Su et al. (2004) found that cost-complexity pruning as developed in standard CART (Breiman et al., 1984) methodology was inferior to several other pruning criteria. They found the BIC (Schwarz, 1978) and RIC (Shi and Tsai, 2002) to give the best results. This issue has also been addressed in the noisy square wave simulations below. For routine tree selection BIC is recommended.

After these first simulations a more complex model was used, i.e.

$$y_t = \mu_{r_i} + \epsilon_t$$

where $\mu_{r_i}$ = the mean of regime $r_i$; $i = 1, \ldots, 5$ and $\epsilon_t$ = noise terms drawn from an N(0,1) distribution. In all simulations $\mu_{r_i} = 0$ for $i = 1, 3, 5$ and $\mu_{r_4} = -\mu_{r_2}$. When ART was used to test the series the value of $\mu_{r_2}$ started at 2 standard deviations and was decremented to 0.05 in steps of 0.05. Because of the large amount of computation required when BPP was used the value of $\mu_{r_2}$ was decremented to 0.1 in steps of 0.1.
In essence the resultant series are square waves with an amplitude of break size and gaussian noise of constant variance imposed on them. The noisy square wave simulations were tested by BPP and by ART with $\alpha = 0$, tree selection by BIC, and tree selection by $T$-fold cross-validation where $T$ is the length of the series. Consistent with Su et.al. we found BIC was a good choice. $T$-fold cross-validation was better for short series but is computationally expensive.

The results from these simulations using ART are presented in Figure (2). BPP results are presented in Figure (3). BPP is computationally expensive. It took 633,918 seconds of CPU time on an 750Mhz UltraSPARC III to generate and analyze 2000 series of length 720 with BPP compared to 1368 seconds on an 1.5Ghz UltraSPARC IIIi processor to generate and analyze 40,000 series of the same length with ART.

When the break-size is large and the regime size is long ART consistently finds the real breaks and generating few spurious candidate breaks without needing to prune $T_{\text{max}}$.

ART’s robustness to non-normal noise was tested with series with gamma and geometric noise structures. Our results (not presented here) show only a slight reduction in ART’s ability to correctly locate the break with similar
FIGURE 3. Total number of breaks found by BPP in the noisy square wave simulations.

numbers of spurious breaks reported.

References


Analyzing forest fire scar data

William J. Reed

Abstract: A method of analyzing fire scar data with the aim of estimating historical forest fire frequency is discussed. Finding maximum likelihood estimates involves maximizing a quasi-likelihood with missing data, via the EM algorithm.

Keywords: Forest fire; fire scars; overdispersion; quasi-likelihood; EM algorithm.

1 Introduction.

Fire plays an important role in many natural ecosystems. In order to implement so-called ‘ecosystem management’ of conserved natural areas of forest and other wild land, in which human activity is tailored to mimic natural processes, it is necessary to have estimates of the rates of naturally occurring fires and to have information concerning the size distribution of fires. For stand-replacement fires, which occur in the boreal forests of N. America, it is possible to estimate historical fire rates from the current age distribution of (even-aged) stands (Reed et al. 1998; Reed, 2000). However in other areas, where surface fires rather than crown fires predominate, uneven-aged stands will occur, and these methods cannot be used. In such areas composite fire interval charts have been constructed. These comprise fire event chronologies based on individual trees with multiple scars or on plots with several trees registering scars.

Because the same fire may register scars on several trees, while at the same time a given tree can contain scars from several fires, the statistical analysis of such data is not straightforward and requires careful statistical modelling.

Fig. 1 shows a very simplified (artificial) composite fire interval chart. Fires are represented by the vertical dotted lines. The fires are of varying extent, thereby registering scars (represented by crosses) on only some of the sample trees (or plots) represented by horizontal lines. Note that not all sample trees are present throughout the whole period.

Methods.

It is assumed that the study area is homogeneous with respect to its vulnerability to fire. If this is clearly not the case then the study area can first be partitioned into homogeneous sub-areas, which can be analyzed separately. For the purpose of analysis, the period of observations is divided into
FIGURE 1. Typical form of fire-scar data. The horizontal lines represent sample trees and the crosses registered scars. Crosses which align vertically correspond to scars registered by the same fire. The total time is divided into distinct epochs (labelled E1, E2, etc., during each of which the number of sample trees present is constant.

non-overlapping epochs, during which the same number of sample trees are vulnerable to fire. For the data in the figure there are six such epochs, labelled ‘E1’, ‘E2’, etc. at the top of the figure.

Let $\lambda$ denote the hazard rate for the occurrence of a fire, which registers a scar somewhere in the study area (the area-wide hazard). Now given such a fire occurs let $p$ denote the conditional probability that it leaves a scar on a given sample tree, so that $\theta = \lambda p$ is the hazard rate for a scar being registered on a given tree (the local hazard). The main interest in the paper is in estimating $\theta$ and its reciprocal, known in the fire ecology literature as the fire return interval.

Of course fires spread spatially, so the events of scars being registered on distinct sample trees are not independent. Given that a fire occurs some-
where, the number of sample trees registering a scar will have a distribution exhibiting over-dispersion relative to a binomial distribution. One possibility is to use a a parametric model such as the beta-binomial, to capture the over-dispersion. Alternatively one can use quasi-likelihood methods for a general overdispersed exponential family model. This is the approach pursued in the paper. A novelty arises in that zero values of the overdispersed binomial distribution are unobserved. This happens when fires occur, but register no scars on the sample trees. To deal with these “missing data” one can use the E-M algorithm to maximize the quasi likelihood.

<table>
<thead>
<tr>
<th>Epoch</th>
<th>No. of objects</th>
<th>No. of fires</th>
<th>Nos. of scars</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>53</td>
<td>3</td>
<td>51, 1, 1</td>
<td>**</td>
</tr>
<tr>
<td>4</td>
<td>59</td>
<td>2</td>
<td>2, 1</td>
<td>.56</td>
</tr>
<tr>
<td>10</td>
<td>67</td>
<td>3</td>
<td>1, 1, 2</td>
<td>.48</td>
</tr>
<tr>
<td>11</td>
<td>68</td>
<td>4</td>
<td>1, 1, 1, 5</td>
<td>.10</td>
</tr>
<tr>
<td>12</td>
<td>69</td>
<td>2</td>
<td>1, 44</td>
<td>**</td>
</tr>
<tr>
<td>13</td>
<td>70</td>
<td>7</td>
<td>5, 2, 1, 1, 2, 1, 57</td>
<td>**</td>
</tr>
<tr>
<td>14</td>
<td>71</td>
<td>5</td>
<td>8, 1, 29, 1, 64</td>
<td>**</td>
</tr>
<tr>
<td>15</td>
<td>72</td>
<td>10</td>
<td>1, 3, 23, 2, 66, 1, 9, 1, 1, 7</td>
<td>**</td>
</tr>
<tr>
<td>16</td>
<td>71</td>
<td>8</td>
<td>16, 8, 12, 7, 36, 2, 1, 60</td>
<td>**</td>
</tr>
<tr>
<td>17</td>
<td>70</td>
<td>6</td>
<td>2, 3, 22, 31, 12, 51</td>
<td>**</td>
</tr>
<tr>
<td>18</td>
<td>68</td>
<td>3</td>
<td>1, 3, 32</td>
<td>**</td>
</tr>
<tr>
<td>19</td>
<td>66</td>
<td>10</td>
<td>27, 2, 47, 1, 5, 3, 21, 23, 1, 35</td>
<td>**</td>
</tr>
<tr>
<td>20</td>
<td>65</td>
<td>5</td>
<td>11, 6, 54, 1, 47</td>
<td>**</td>
</tr>
<tr>
<td>24</td>
<td>56</td>
<td>3</td>
<td>5, 4, 7</td>
<td>.62</td>
</tr>
<tr>
<td>25</td>
<td>53</td>
<td>2</td>
<td>2, 21</td>
<td>**</td>
</tr>
<tr>
<td>29</td>
<td>38</td>
<td>2</td>
<td>3, 16</td>
<td>.0006</td>
</tr>
<tr>
<td>34</td>
<td>12</td>
<td>3</td>
<td>2, 1, 5</td>
<td>.14</td>
</tr>
</tbody>
</table>

The quasi-likelihood function is derived as follows. In epoch \( j = 1, 2, \ldots, J \)
let
\( \tau_j \) denote the duration of epoch;
\( N_j \) denote number of trees in the sample vulnerable to scarring;
\( n_j \) denote number of distinct dates at which fires were recorded;
\( t_{j,r} (r = 1, \ldots, n_j) \) denote the times (since start of epoch) at which fires were recorded;
\( x_{j,r} (r = 1, \ldots, n_j) \) denote number of scars registered for fire \( i \) and
\( x_j = \sum_{r=1}^{n_j} x_{j,r} \) denote total number of scars registered in epoch.
Also let
$m_j$ denote the number of dates when fires which registered no scars on the
sample trees were ignited in epoch $j$.

N.B the $m_j$ are unobserved variables.

The total number of fires ignited in the area over the total time $T$ (since
start of first epoch) is

$$\sum_{j=1}^{J} n_j + m_j = n_+ + m.$$ 

We wish to estimate model parameters $\lambda$ (area-wide hazard of burning)
and $p$ (the conditional probability of a fire at any location given a fire
somewhere) and from them the local hazard of burning $\theta = \lambda p$.
The “full-data” likelihood is the probability of such observations (including
“missing values” $m_j$, $j = 1, \ldots, J$ - the numbers of fires leaving no scars).
Now if scars were registered independently of one another on sample ob-
jects, given that a fire has ignited, the distribution of the number of scars
registered ($X_{j,r}$) for a given fire ($r$) in epoch $j$ would be

$$\text{Bin}(N_j, p)$$

Clearly because of the fact that fires spread spatially the assumption of
independence is highly questionable. Indeed the results of a binomial dis-
persion test (for Dugout region of Blue Mountains, E. Oregon) shown in
Table 1 confirm the invalidity of the binomial model.

To allow for contagion in the spread of fires consider instead an overdis-
persed binomial model with pmf $f(x; N, p, \phi)$ where $\phi$ is overdispersion
parameter. So that (with $q = 1 - p$)

$$\ln f(x; N, p, \phi) = \frac{x \log(p/q) + \log q^{N}}{\phi} + c(\phi, x).$$

Under these assumptions the (full data) quasi-likelihood is of the form

$$L(\lambda, p; \phi) = k \lambda^m + n e^{-\lambda T} \prod_{j=1}^{J} f(0; N_j, p, \phi)^{m_j} \prod_{r=1}^{n_j} f(x_{j,r}; N_j, p, \phi)$$

($NB.$ this includes the “missing data” $m_j$, $j = 1, \ldots, J$). The log quasi-
likelihood is

$$Q(\lambda, p; \phi) = c + (n_+ + m_+ \log \lambda - \lambda T$$

$$+ \frac{1}{\phi} \left( \sum N_j m_j + \sum N_j n_j - x_+ \right) \log(1 - p) + x_+ \log p \right)$$
MLEs of $\lambda$ and $p$ can be found by maximizing the quasi-likelihood over $\lambda$ and $p$ using the EM algorithm; and an estimate of the overdispersion parameter $\phi$ can be found using the method of moments.

It turns out that EM can be done analytically. The M-step and E-step together yield the following (coupled) recurrence equations relating the estimates at $i+1$ step, with those at $i$ step:

$$\hat{\lambda}^{(i+1)} = \frac{n + \hat{\lambda}^{(i)} \sum (1 - \hat{p}^{(i)}) N_j \tau_j}{T}$$

$$\hat{p}^{(i+1)} = \frac{x_j}{\hat{\lambda}^{(i)} \sum N_j \tau_j (1 - \hat{p}^{(i)}) N_j + \sum N_j n_j}.$$ 

The coupled system converges to an equilibrium which yields the MLEs $\hat{\lambda}$ and $\hat{p} = 1 - \hat{q}$

$$\hat{\lambda} = \frac{n}{T - \sum \hat{q} N_j \tau_j}$$

$$1 - \hat{q} = \frac{x_j (T - \sum \hat{q} N_j \tau_j)}{n \sum N_j \tau_j \hat{q} N_j + \sum N_j n_j (T - \sum \tau_j \hat{q}^2 N_j)}$$

NB. The second equation has to be solved numerically.

2 Results and conclusions.

The method was applied to data collected in the Blue Mountains of eastern Oregon, U.S.A. by E-K. Heyerdahl (Heyerdahl 1997 and Heyerdahl et al. 2000). We use four sites: Tucannon and Imnaha (both of which have north and south facing hillslopes), Baker (northeast facing hillslopes) and Dugout (west facing hillslope).

Each site was divided into cells each approximately 25 ha. A one ha plot was placed in the center of each cell. A fire event chronology was contracted from fire scars and tree ages for each one ha plot. The south-facing and north facing parts of the Tucannon and Imnaha sites are treated separately for analysis making six study areas in all. Table 3 gives estimates of the fire interval in the six areas.

The confidence intervals for the fire interval are quite wide, especially for the wetter north-facing sites. This poor precision is typical in the analysis of fire scar data occurring because of the relative scarcity of fires and of registered scars. Increasing the number of sampled trees could improve the precision of the estimate of the parameter $p$, but would likely have little effect on the precision of the estimate of $\lambda$ (because scars from already identified fires would likely predominate over scars from previously unidentified fires).
TABLE 2. Estimates of the fire interval (reciprocal of local hazard) for sites in the Blue Mountains. Note that for all confidence intervals if the lower limit is negative it is reported as zero.

<table>
<thead>
<tr>
<th>Site (aspect)</th>
<th>MLE of FI (years)</th>
<th>standard error</th>
<th>estimated dispersion, $\hat{\phi}$</th>
<th>95% CI for FI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tucannon (N)</td>
<td>183.50</td>
<td>102.3</td>
<td>6.92</td>
<td>0 - 384.0</td>
</tr>
<tr>
<td>Tucannon (S)</td>
<td>42.2</td>
<td>8.8</td>
<td>8.05</td>
<td>24.9 - 59.4</td>
</tr>
<tr>
<td>Imnaha (N)</td>
<td>118.2</td>
<td>79.8</td>
<td>21.16</td>
<td>0 - 274.6</td>
</tr>
<tr>
<td>Imnaha (S)</td>
<td>34.2</td>
<td>13.23</td>
<td>57.32</td>
<td>8.2 - 60.1</td>
</tr>
<tr>
<td>Baker (NE)</td>
<td>23.0</td>
<td>3.78</td>
<td>9.84</td>
<td>15.6 - 30.4</td>
</tr>
<tr>
<td>Dugout (W)</td>
<td>21.7</td>
<td>3.65</td>
<td>28.06</td>
<td>14.5 - 28.8</td>
</tr>
</tbody>
</table>

References


Modelling zero inflation of compositional data

Michael Salter-Townshend and John Haslett

1 Dept. of Statistics, School of Computer Science and Statistics, Trinity College Dublin, Ireland

Abstract: In their paper on a computationally intensive MCMC method for palaeo-climate reconstruction from multivariate pollen counts, Haslett et al (2006) encountered substantial zero inflation; a database of 7815 14-dimensional counts contained 36% zeroes. This paper introduces a new model for such zero-inflated multivariate count data in a spatial context. Specifically, we show that a spatial Gaussian process can be conveniently used to create a zero-inflated mixture of the Multinomial.

Keywords: Bayesian inference; prior; zero-inflation; overdispersion; compositional data.

1 Introduction

Haslett et al (2006) reported on the reconstruction, from fossil pollen, of the palaeoclimate at Glendalough in Ireland. The essential science is that pollen found in lake sediment reflects ancient vegetation, which in turn reflects the ancient climate. Thus changes in the pollen composition with sediment depth reflect changes in the climate with past time. This paper concerns one aspect of the statistical modelling used to describe the variation in pollen composition; such modelling is an essential part of a computationally-intensive methodology whose main contribution is a focus on the modelling of the uncertainties in climate reconstruction. In particular this paper focusses on Bayesian approaches to zero-inflation in multivariate count data, in the context of the non-parametric smoothing used in palaeoclimate reconstruction. We model two sources of extra zeroes; structural zeroes arising where the climate is unsuitable for a particular plant to grow and zero-inflation of the counts data where there is a potential presence.

The composition of the pollen in a sample is reflected in a multivariate count vector $y$ of the distinguishable taxa. The focus is on the observed proportions $\hat{p} = y/n$. The computationally intensive methods discussed in Haslett at al (2006) were illustrated with $y$ of dimension 14; in Huntley et al (1993) the dimensionality is 28; the total count is typically about $n = 400$. Variation in $y$ is two-fold: random and systematic; this is illustrated in Figure 1 and Figure 2. Here we show aspects of modern data on composition
at 7815 locations in the Northern Hemisphere at which the modern climate \( c \) is known; here it is represented by two aspects - \( GDD_5 \) and \( MT CO \); these are, respectively, the Growing Degrees Days above 5°C being a measure of the length of the growing season, and \( MT CO \), the mean temperature of the coldest month. Such modern data are used as training data to fit such surfaces. Haslett et al (2006) used these two climate dimensions; in Huntley et al (1993) three climate dimensions are used.

In Figure 1 systematic variation of \( Betula \) is illustrated via a non-parametric smooth in (here) two-dimensional climate. This reflects the fact that each taxon has one or more ‘preferred’ climates in which it is most productive. At about \( GDD_5 = 1500 \) and \( MT CO = -7.5 \) \( Betula \) is particularly productive and is often found in ancient pollen. Such a smooth is referred to as a ‘response surface’.

In Figure 2 the random variation in the components corresponding to \( Betula \) (Birch) and \( Corylus \) (Hazel) are presented for a small region of two dimensional climate space \( GDD_5 = 1500 \) and \( MT CO = -7.5 \). Zero inflation is apparent in both. \( Betula \) often comprises 40% of the composition in this small region of climate space; yet it is frequently completely absent; by contrast, even when \( Corylus \) is present (which is rare for this climate) it comprises only a very small percentage of the composition. Figure 3 provides a schematic on which we rely for this paper; it is a simpler version of Figure 1. Climate is one dimensional and there are but two taxa; the compositional vector \( (\hat{p}(c), 1 - \hat{p}(c)) \) has two components, of which we show only the first; the second component may be thought of as ‘other’ taxa. The counts, at certain known climates \( c_j \), are shown as proportions \( \hat{p}(c_j) \); these correspond to the modern training data for which both \( p \) and \( c \) are known. In Figure 3, there are structural zeroes to the left of \( c = 400 \). Zero-inflation occurs randomly at all points due to the lower curve \( q \). We expand below on \( r \) and \( q \).

The reconstruction task may be stated simply in Bayesian terms: given a forward model for the random variation in pollen counts \( y(c) \sim \pi(p(c), n) \) (built on data such as in Figure 1, where \( E[y(c)/n] = p(c) \)), construct the posterior distribution \( \pi(c|\hat{p}_0) \) of climate \( c \), given additionally one or more count vectors \( y_0 \) taken from sediment, and thus corresponding to unknown climate. Several issues of smoothness arise: \( p(c) \) varies smoothly with climate and climate \( c(s, t) \) itself varies smoothly with space \( s \) and time \( t \). Haslett et al (2006) approach such smoothness non-parametrically in a Bayesian fashion via Gaussian process priors. The computational challenge lies in the high dimensionality of \( p \) and \( c \) and in the number of points in space and time at which reconstructions are sought. Zero-inflation in the forward model is an added complication for which we seek a parsimonious approach; this, in the context of non-parametric smoothing, is the contribution of this paper. More specifically we discuss a richer description
of random variability than is modelled through $\pi(p(c), n)$; yet it remains parsimonious.

Random variation - at the stage of sampling and counting - in a single taxon such as Betula is most simply modelled via the Binomial distribution. The left side of Figure 2 shows however that there can be substantial zero-inflation with respect to the Binomial - although this instance is extreme. Conjugate mixtures such as the Beta-Binomial have been used as a computationally convenient approach to over-dispersion; the multivariate equivalent - a Dirichlet mixture of Multinomials - was used in Haslett et al (2006). Below we discuss models focussed more specifically on zero-inflation. Systematic variation with climate $c$ is modelled via the distribution of the vector function $\hat{p}(c) = y(c)/n$. Non-parametric smoothing of zero-inflated compositional data is thus the focus of this paper.
2 Zero-inflated Statistical Models in Climate Space

In this paper our focus is on the essential modelling issues; we use much more simplified data to present the ideas. The notation reflects the simpler Figure 3; generalisation to higher dimensions of \( y \) and \( c \) is discussed subsequently, but briefly. We refer to climate as ‘climate space’ and describe the distribution of \( y(c) \) as varying spatially in a smooth fashion, even though in this section \( c \) has only one dimension. Given univariate data \( \hat{p}(c) = y(c)/n \), we seek the posterior distribution of smooth univariate functions \( (r(c), q(c)) \) that parameterise the random variation in \( y(c) \) as discussed below (see Section 2.1); but we show that a very rich description is available with one such function. Our eventual interest is in the posterior of an unknown climate given a count vector and known climates with count vectors. In the usual Bayesian paradigm we write

\[
\pi(c|\hat{p}(c)) \propto \pi(\hat{p}(c)|r(c), q(c)) \times \pi(r(c), q(c)) \times \pi(c)
\]

We sample from the posterior probability for the proportions using a Metropolis-Hastings Markov Chain Monte Carlo algorithm.

More typically \( (r(c), q(c)) \) will be evaluated only on a discrete grid \( c_g \) in climate space, being thus defined on \( c_g \), with implications, of course, in the technical definition of ‘smoothness’. We discuss options below.

2.1 Zero-inflation

As motivated by Ridout et al (1998), we propose the mixture model for the univariate pollen count variable \( y \):

\[
Y = 0 \text{ with probability } 1 - q
\]

\[
Y \sim Bin(n, r) \text{ with probability } q
\]

where \( r = r(c) \) and \( q = q(c) \) are functions of climate; \( q(c) \) is the probability the taxon is potentially present at climate \( c \); when present, the expected proportion is \( r(c) \). Note that \( E[\hat{p}] = p = r \times q \). The corresponding non-zero-inflated model is \( y \sim Bin(n, p) \). The posteriors for \( c \) are thus

\[
\pi_{c,q}(c|y, n) = I_{y=0}(1 - q(c)) + q(c)^y(1 - r(c))^{(n-y)}
\]

\[
\times \pi(r(c), q(c)) \times \pi(c)
\]

and

\[
\pi_p(c|y, n) = (\binom{n}{y})p(c)^y(1 - p(c))^{(n-y)} \times \pi(r(c), q(c)) \times \pi(c)
\]

Figure 4 shows the \( r(c), q(c) \) and \( p(c) \) functions. Figure 5 shows the corresponding posteriors of \( c \) given non-zero and zero counts under these two models.
FIGURE 3. Toy problem functions and data when there are two unrelated functions of climate, \( r(c) \) and \( q(c) \). The top plot shows the response curve \( r(c) \). \( r(c) \) is passed to the Binomial to generate counts data (plotted as proportions) when present. The lower curve is \( q(c) \), the probability the taxon is potentially present.

For \( y \neq 0 \) these are effectively equivalent. However, when \( y = 0 \) the first model tells a richer story. For example at \( c = 7 \), \( q \) is low but \( r \) is high; the probability of absence is high and therefore the likelihood of a zero count under the \((r(c), q(c))\) model is high. The posteriors for the zero-inflated and non zero-inflated \( \text{Bin}(n, p) \) models are shown for comparison. The prior on \( c \) is Gaussian.

FIGURE 4. Functions \( r(c) \), \( q(c) \) and \( p(c) = r(c) \times q(c) \) used here for illustrative purposes. In (a) they are unrelated and in (b) they are almost coincident. \( r(c) \) is the solid line, \( q(c) \) is dashed and \( p(c) = r(c) \times q(c) \) is dotted.

A simpler model has \( q(c) = r(c) \). For example, at a particular climate the probability that the taxon is present might only be 0.3; but further, when present, it comprises only 30% of the pollen, on average. Under this model,
the environmental pressure exerted by climate influences in a similar fashion both the probability of presence and the expected proportion, if present. Figure 4(b) again shows the $r(c)$, $q(c)$ and $p(c)$ functions with posteriors of $c$ given $y = 5$ and $y = 0$. The likelihoods and therefore posteriors for both non-zero and zero counts are now closer under the two models. The posterior for a zero count is somewhat different as zero inflation still needs to be addressed in the $p(c)$ model. More general models of linking $q(c)$ and $r(c)$ are possible.

2.2 Gaussian processes

A Bayesian implementation requires priors $\pi(r(c))$ and $\pi(q(c))$. In this paper we adopt the conditional Gaussian priors used in Haslett et al (2006). Specifically at a single point in climate space, we write, for $0 < r < 1$, $\pi(r) \propto \phi((r - \alpha_r)/\beta_r)$, where $\phi$ is the standard normal density. We can represent this as $r = z_r$, conditionally on $0 < z_r < 1$, where $z_r \sim \mathcal{N}(\alpha_r, \beta_r^2)$ unconditionally. We can similarly represent $\pi(q)$.

To ensure smoothness in discrete climate space, we represent the joint prior for $\{r(c); c \in c_g\}$ as that of a realisation of a discrete Gaussian process $\{z_r(c); c \in c_g\}$ constrained such that all elements of $\{z_r(c); c \in c_g\}$ lie in the interval $(0, 1)$. In this paper, the processes used to provide these priors are first order auto-regressions.
2.3 Structural Zeroes

We use such processes to introduce another aspect of pollen response to climate. Specifically, if we constrain such a process such that all elements of $z_r(c); c \in c_g$ are less than unity, writing $r(c) = z_r(c)$ when $0 < z_r(c) < 1$ and $r(c) = 0$ when $z_r(c) \leq 0$, then we may find 'regions of climate space' where $r(c) = 0$. Such a process is not zero-inflated at such a point; it is simply that $P(y(c) = 0) = 1$. But such points can be common.

2.4 Remarks

Other link functions, such as $\text{logit}(r(c)) = z_r(c)$ can be used; we do not discuss such issues here. Note that we can envisage many variations, such as different convolutions, of a common $z(c)$ for $r(c)$ and $q(c)$. The key here is that a single underlying process can be used to generate many models of spatial zero inflation. It is this that renders the model computationally attractive in the context of high dimension palaeo climate reconstruction.

3 Results

The model for zero inflation which equates expected proportions with probabilities of presence performs well in reconstructing response curves from data generated using a single underlying process. (e.g. $q(c) = r(c)$). Figure 6(a) shows the random function $r(c)$ and the data generated from it and $q(c) = r(c)$ along with the modal and 95% highest posterior density regions for the reconstructed curve. However, for data generated using two distinct processes, (i.e. when $r(c)$ and $q(c)$ are unrelated functions), the model fails due to zero inflation of the data where the response $r(c)$ is high. An extreme case of this is illustrated in Figure 6(b), where $r(c)$ and $q(c)$ are very different.

4 Conclusions

The model for zero-inflation through a presence probability based on the same process as the expected proportions when present is readily extended to the Multinomial framework in more than one climate dimension. It is faster than the model with separately specified $r(c)$ and $q(c)$, but in order to be applicable, it requires that the response and presence functions on climate arise from a single underlying process. For our pollen data this seems an intuitive assumption. A plant will be least likely to be absent at its most preferred climate. The scenario depicted on the left in Figure 2 provides a counter example; however this situation is rare. Much more common is the sort of variation shown on the right of Figure 2.
FIGURE 6. Posterior modes and 95% highest posterior density regions for $r(c)$. Counts data are plotted as proportions. In (a) $q(c) = r(c)$; in (b) the two functions $r(c)$ and $q(c)$ are unrelated. The model based on a common $r(c)$ and $q(c)$ copes well with the first scenario but struggles with the second.

Additional overdispersion (including zero inflation) may be modelled in an inexpensive manner through mixing of the likelihood with a suitable conjugate distribution. For example the Dirichlet mixture of the Multinomial used in Haslett et al (2006) introduces just one extra variable. We have used $q(c) = r(c)$ as the relationship between the probability of absence and the expected proportion when present; more general models of linking $q(c)$ and $r(c)$ are possible. We propose to apply this richer model to multiple taxa in a higher dimensional climate space.

Acknowledgments: Michael Salter-Townshend and Prof. John Haslett have been supported by Enterprise Ireland grant SC/2001/171 and SFI grant 04/BR/M0049. We acknowledge the assistance of Dr. Andrew Parnell, supported by SFI grant 05/RFP/MAT0044.

References


456 Zero inflation of compositional data

Extreme value modelling of reactor risk

Carl Scarrott¹, Granville Tunnicliffe-Wilson² and Jon Tawn²

¹ Maths and Stats Dept, Uni of Canterbury, Christchurch, New Zealand,
carl.scarrott@canterbury.ac.nz
² Maths and Stats Dept, Uni of Lancaster, Lancaster, LA1 4YF, UK.

Abstract: A statistical approach to modelling the risk of critical temperature exceedances within a nuclear reactor is presented. This paper focuses on extreme value modelling of the upper tail of the risk distribution which is of most concern. A model based bootstrap is developed to quantify the uncertainty in the risk predictions. A validation of the methodology for use on other reactors is discussed.

Keywords: Extreme values; spatial random effects model; bootstrap.

1 Introduction

Reliable estimation of the risk of critical temperature exceedance in nuclear reactors is an essential requirement towards safe operation and compliance, as is an understanding of the uncertainties associated with any estimates. Physical (deterministic) modelling of reactor cores is challenging, due to the complexity of the processes involved and their inherent stochastic nature. A statistical approach to modelling Magnox reactor core temperatures has been developed by Logsdon et al. (2000), Scarrott and Tunnicliffe-Wilson (2001) and Scarrott (2003). This paper focuses on extreme value modelling of the residual risk distribution to estimate the expected number of critical temperature exceedances and bootstrapping for uncertainty estimation.

1.1 Background

Early Magnox reactor core designs take a sample of fuel channel gas outlet temperature measurements (CGOT’s) at fixed locations to monitor and control the fission rate across the reactor core. The 6156 fuel channel outlets form a spatial lattice over a roughly circular region, see Figure 1 for layout of the 2 Wylfa reactors (located on Anglesey, Wales). The Wylfa reactors are the only Magnox reactors in which snapshots (scan’s) of the population of CGOT’s are periodically taken, in addition to the sample of continuously monitored control measurements.

A key challenges for quantifying risk in these reactors is that the sample of CGOT measurements used for reactor control may exhibit a different
distribution to the unobserved measurements. In particular, no measured temperature should exceed the reactor operating limit. However, some unobserved CGOT’s may exceed this limit, see Logsdon et al. (2000) for further discussion. This “control effect” means that standard statistical inference is not possible, as the distribution of sample measurements is not representative of the population of CGOT’s, particularly in the upper tail which is of most concern.

![Figure 1](image1)

**FIGURE 1.** Scan CGOT measurements in °C for Wylfa R1 (reactor 1) in March 97. Missing values are white.

### 2 Risk Methodology

A statistical model for predicting the unobserved temperatures using the observed control measurements is briefly outlined in the following section. Figure 2 is a transect across the spatial temperature surface, depicting the risk estimation procedure developed by Logsdon et al. (2000). The crosses are the sample of control measurements, with the predicted temperature surface from the statistical model shown by solid line. The remaining uncertainty impacting the exceedance risk is the residual variation about the predicted temperature surface. The exceedance probability for each channel is determined by the survivor function of the observed residual distribution.

Extreme value modelling of the upper tail of the residual distribution allows reliable extrapolation of the exceedance probabilities beyond the observed
range. The model residuals are assumed spatially homogeneous, independent of each other and all explanatory variables. A further important assumption is that the observed residual distribution using a sub-sample of CGOT’s is representative of the unobserved channels. Hence, the residuals are pooled over all observed channels to quantify the “residual risk distribution” at all channels. The exceedance probabilities are summed over all channels to give the expected number of exceedances for a range of critical thresholds, termed the “risk predictions”.

![Diagram of temperature exceedance risk estimation procedure](image)

**FIGURE 2. Schematic of temperature exceedance risk estimation procedure.**

### 2.1 Statistical Model

A linear mixed model was developed by Scarrott and Tunnicliffe-Wilson (2001) and Scarrott (2003) to predict the population of CGOT’s. Fixed effects encapsulate spatial temperature variation due to reactor conditions: fuel irradiation, geometry, etc. These deterministic components cause the apparent regular spatial structure of the CGOT’s in Figure 1, e.g. checkerboard, east-west bands, cooler outer ring. Random effects model the slowly varying spatial variation (including control effect) and other stochastic spatially structured variation due to artifacts of the measurement process.

The statistical model is developed using data from the Wylfa reactors, for which snapshots of the CGOT’s from all channels are available (excluding a small number of missing measurements considered missing at random). The model parameters are estimated using two datasets:

- **full model** - using all valid CGOT measurements, as a benchmark
• **3x3 model** - using sample of CGOT measurements commensurate with number/spread of control measurements in other reactors. Every third channel in the $(x, y)$ directions is used, giving a “3x3 sub-grid”.

The 3x3 model allows evaluation of the potential performance of the methodology in reactors where only the sample of control measurements are taken. Leave one out cross-validation predictions are used in both models, so as to provide a realistic measure of predictive performance.

### 2.2 Extreme Value Modelling

The probability of critical temperature exceedance needs to be reliably estimated past the observed range of the empirical residual distribution, see Figure 2. This extrapolation is particularly challenging in reactors where only the sub-grid of residuals are observable.

The generalised Pareto distribution (GPD) for the upper tail is an asymptotically justified model for the excesses over a “suitably high threshold $u$” for a sequence of IID random variables. The GPD has seen wide application in many fields, see Coles (2001) for discussion. If the random variable $R$ represents a possible residual, the excess $Y = R - u$ over the threshold $u$ follows a GPD parameterised by scale $\sigma_u$ and shape $\xi$, with cdf:

$$Pr(Y < y | Y > 0) = 1 - \left(1 + \xi \left( \frac{y}{\sigma_u} \right) \right)^{-1/\xi}$$

where $x_+ = \max(x, 0)$, which explicitly conditions on exceeding the threshold $u$. The shape parameter $\xi$ determines the tail behaviour; $\xi = 0$ equates to an exponential distribution ($0 < y < \infty$), $\xi < 0$ a tail with finite support ($0 < y < -\sigma_u/\xi$) and $\xi > 0$ a heavier tail than exponential ($0 < y < \infty$).

If $u$ is a suitably high threshold then the GPD is also an appropriate model for exceedances for any higher threshold $v > u$. The shape parameter for all higher thresholds $v > u$ will be the same, although empirical estimates may vary. However, the scale parameter linearly changes with threshold, hence the reference to $u$ in the symbol $\sigma_u$.

Maximum likelihood estimation is used for GPD parameters, a standard procedure with known statistical properties (Coles, 2001). Various diagnostics (Coles, 2001) were used to determine the threshold, e.g. mean residual life plot, threshold parameter stability plots, see Scarrott (2003) for full details. As with the statistical model above, the GPD is fitted to two datasets:

- **full grid** - all full model cross-validation residuals, as benchmark
- **3x3 on-grid** - 3x3 model cross-validation residuals at simulated control channels (on-grid), to assess performance for other reactors.
The observed residual distribution is assumed representative of all channels. For the full model this is a very reasonable assumption to make. However, for the 3x3 model it leads to a conservative bias in the exceedance probabilities. The variance of the predictions at the observed (on-grid) channels will always be larger than at unobserved (off-grid) channels as cross validation is used to determine the predictions. When predicting at an on-grid location the nearest observation will be at least 3 channels away, whereas for an off-grid location the nearest observation will be closer and so the prediction will have a smaller variance associated with it.

The probability that the CGOT $T_k$ in channel $k$ exceeds the critical temperature $\tau$ is estimated using:

$$Pr(T_k > \tau) = Pr(T_k - \hat{T}_k > \tau - \hat{T}_k) \approx Pr(R > \tau - \hat{T}_k)$$

where $\hat{T}_k$ is the predicted temperature at channel $k$ and $R$ is random variable for observable residuals. The GPD is used to calculate exceedance probabilities when the difference $\tau - \hat{T}_k$ is above the threshold $u$. Exceedance probabilities for differences below the threshold $u$ are estimated from the empirical residual distribution, as these are not critical for risk assessment and can be reliably estimated using the available data (for both full and 3x3 on-grid residuals). The GPD as defined above is conditional on the difference exceeding the threshold $u$. The unconditional probability is obtained by multiplying it by the probability of exceeding the threshold, which is simply estimated using the observed proportion of threshold exceedances.

Given the assumptions about the residuals outlined in Section 2, the expected number of exceedances of a critical temperature can be estimated by summing the exceedance probabilities for each channel. A plot of the expected number against a range of different critical temperatures is termed the risk predictions. The risk predictions can be validated for low critical temperatures using the observed number of exceedances.

### 2.3 Application

Figure 3 is an example risk prediction at one timepoint (state). Risk predictions were produced for 11 states (5 for Wylfa R1 and 6 for Wylfa R2) for which data was available, which all exhibited similar features to those presented. Table 1 gives the GPD parameter estimates. The shape parameter for the full grid and 3x3 on-grid are marginally significantly less than zero (at 95% level) for almost all states, signifying a finite upper bound. The full model predictions closely follow the observed number of exceedances and provides a visually sensible extrapolation. The 3x3 risk prediction, is indicative of potential performance on other reactors. The 3x3 on-grid predictions are pessimistic (higher than full model over range plotted), which is consistent across all states examined. This pessimism is due to cross-validation being used to determine the on-grid residuals, discussed in Section 2.2.
Logsdon et al. (2000) pooled the residuals from a number of states to improve the risk predictions, under the assumption the residuals are homogeneous through time. Analysis of the residuals from the mixed model outlined above showed evidence of heterogeneity over time. In particular, the variance decreased through time which is hypothesised to be due to data quality improvements. The risk predictions appeared to be well estimated using only the data within each state, hence it was deemed unnecessary to use the pooling over states, thereby avoiding the need for assuming homogeneity across states. An analysis of the risk predictions using the residuals pooled over all 11 states lead to no substantive qualitative difference to the risk predictions. Logsdon et al. (2000) also ignored residuals where the predictions were greater than 375°C, to ameliorate any remaining control effect. Extensive exploratory analyses found no evidence for any remaining control effect for the mixed model residuals.

Exploratory analyses and likelihood ratio tests were used to determine if remaining variation in the upper extreme residuals could be explained by various explanatory variables, but no significant associations were found. Various spatial dependence measures were used to detect any dependence in the location and magnitude of the upper extreme residuals, similarly no
substantial dependence was found.

The impact of threshold choice for the GPD on the risk predictions was explored, but no substantive impact was found. Further, the arbitrary choice of sub-grid offset (i.e. 9 possible 3x3 sub-grids to choose from) was found to have no substantial effect on the conclusions.

3 Uncertainty Estimation

Bootstrap tolerance intervals are used to quantify uncertainty due to estimation of the residual risk distribution and GPD parameters. Let \( n_c = 6156 \) be the number of channels, \( n_p \) the number of (on-grid) residuals used to define the residual risk distribution and \( n_u \) the observed number of residuals exceeding the fixed threshold \( u \). In order to simulate a new residual risk distribution, the number of realised residuals above/below the threshold must be determined. The threshold is held fixed, but the number above/below the threshold is a binomial random variable.

The simulation procedure is as follows:

1. simulate a Bin\((n_c, n_u/n_p)\) random variable \( n_u^* \) for the number of simulated residuals over the threshold \( u \), giving \( n_c - n_u \) residuals below the threshold,

2. resample \( n_c - n_u^* \) residuals from the observed residuals below the threshold (with replacement),

3. simulate \( n_u^* \) GPD\( (\hat{\sigma}_u, \hat{\xi}) \) excesses of threshold, add \( u \) to give simulated residuals above threshold,

4. pool resampled residuals from below threshold and simulated residuals from above threshold to get simulated residual risk distribution.

5. obtain risk predictions using simulated risk distribution (original mixed model predictions still used)
The simulation procedure is repeated a number of times (100 simulations are used below), giving a set of surrogate risk predictions from which to assess the sampling distribution. The pointwise [5, 95]% quantiles of the simulated risk predictions at each critical temperature is extracted to quantify the uncertainty in the risk predictions.

Example intervals are presented in Figure 3. The uncertainty clearly increases for extrapolations at higher critical temperatures, as expected. Further, the intervals for the full model are substantially smaller than for the 3x3 on-grid risk predictions due to the smaller set of observed residuals used to determine the latter.

The uncertainty bounds using the mixed model developed by Scarrott (2003) were not substantively smaller than for the simpler kernel smooth predictions of Logsdon et al. (2000). The principle reason is that the main source of uncertainty for the risk prediction extrapolations is due to estimation of the GPD parameters, and the intervals do not account for uncertainty due to estimation of the mixed model CGOT predictions. Methods to provide a more complete picture of the uncertainty due to model choice and parameter estimation is a topic of ongoing research.

4 Summary

An existing methodology developed for assessing the risk of critical temperature exceedance in nuclear reactors has been adapted and extended, using a physically justifiable statistical model for predicting reactor temperatures, and bootstrap methods to quantify uncertainty associated with the risk predictions. The potential performance of the methodology for reactors where only a sample of control measurements are available, has been demonstrated. In this case, the risk predictions have been shown to have a consistent conservative bias, but are well estimated and reliable.

Acknowledgments: We gratefully acknowledge the support of Magnox Electric PLC. Financial support was provided by their Industrial Management Committee contract IMC/RPS/5003(a) and EPSRC grant GR/M13848. Some of this work was completed at the Centre for Ecology and Hydrology.

References


A comparative study of false discovery rate procedures in R

Michael G. Schimek\textsuperscript{1,2} and Tomáš Pavlík\textsuperscript{2}

\textsuperscript{1} Medical University of Graz, Institute for Medical Informatics, Statistics and Documentation, Auenbruggerplatz 2, 8036 Graz, Austria
\textsuperscript{2} Masaryk University, Department of Applied Mathematics and Institute of Biostatistics and Analyses, 66295 Brno, Czech Republic

Abstract: Statistical procedures for the identification of differentially expressed genes in microarray studies involve a serious multiple comparison problem as we have to carry out as many hypothesis testings as the number of candidate genes in the experiment. This paper is concerned with the recent type I error control concept of the false discovery rate (FDR), for which an increasing number of competing estimates is available in R. However, there is little comparative evidence. Based on the most popular modified $t$-statistic independent as well as co-regulated gene expression measurements are compared in a simulation study. Finally conclusions are drawn for practical microarray analysis.

Keywords: Correlation; error control; false discovery rate; microarray; multiple testing.

1 Introduction

The microarray technology developed in recent years allows for measuring expression levels of thousands of genes simultaneously. In most microarray experiments comparative measurements are taken from subjects belonging to pre-specified groups. Statistical procedures to identify differentially expressed genes involve a serious multiple comparison problem as we have to carry out as many hypothesis testings as the number of candidate genes in the experiment. If we apply the usual type I error rate $\alpha$ in each testing, then the probability for rejecting any truly null hypothesis will greatly exceed the intended overall $\alpha$ level. In this paper we consider the false discovery rate (abb. FDR) approach for type I error control. Our focus is on recent permutation-based FDR estimation concepts and their realization in the statistical environment R. The motivation of this paper is a lack of comparative evidence. The so-called modified $t$-statistic is studied in combination with several FDR procedures and typical parameter settings, including the use of a tuning constant in the estimation of the pooled variance required for the test statistic. In addition to the situation of independent gene expression measurements the more realistic case of partial
co-regulation (i.e. some expression readings are correlated) is considered in a comparative simulation study. Finally conclusions are drawn for the practice of testing in experimental microarray data.

2 The test statistic

Let us assume a microarray study with two-group design, \( m \) samples belonging to group one and \( n \) samples belonging to group two. Further let us have \( g \) genes and \( (m + n) \ll g \). Let us denote \( \bar{x}_{1} \) and \( \bar{x}_{2} \) the mean gene expression in group one and two, and \( s_{i} \) the pooled standard deviation for gene \( i \). The modified \( t \)-statistic for gene \( i \) is then

\[
    t_{i} = \frac{(\bar{x}_{2} - \bar{x}_{1})}{(s_{i} + s_{0})},
\]

where \( s_{0} \) is a tuning constant, called 'fudge factor' in Tusher, Tibshirani, and Chu (2001) and in the SAM (Significance Analysis of Microarrays; Chu et al., 2005) package. Its specification can be crucial for the number of identified differentially expressed ('significant') genes. For \( s_{0} = 0 \) we obtain the standard \( t \)-statistic for pooled variances. A nonparametric alternative would be the Wilcoxon rank-sum statistic. It is not considered here because Schimek and Pavlík (2006) have provided evidence that the \( t \)-statistic gives better results in most instances.

3 The false discovery rate

The motivation of the false discovery rate is to identify as many differentially expressed genes as possible while incurring a relative low proportion of false positives. The first definition of the \( FDR \) was introduced by Benjamini and Hochberg (1995). Let \( V \) be the number of false positives and \( R \) be the number of overall rejected hypotheses in a microarray experiment. Then the \( FDR \) can be defined as the expectation of the ratio of \( V \) and \( R \) with a supplement to account for the possibility of \( R = 0 \),

\[
    FDR = \mathbb{E} \left( \frac{V}{R} 1_{\{R>0\}} \right).
\]

The advantage of this original formulation is that it also accounts for the dependence between \( V \) and \( R \).

In microarray experiments the situation of \( R = 0 \) is rare so our main aim is to estimate \( FDR = \mathbb{E}(V/R) \). When considering many parameters (usually we test thousands of genes simultaneously), according to Storey and Tibshirani (2003) it can be shown that

\[
    FDR = \mathbb{E} \left( \frac{V}{R} \right) \approx \frac{\mathbb{E}(V)}{\mathbb{E}(R)}.
\]

This approximation is easier to estimate and thus, often without mentioning it, adopted in most implementations. For a discussion of related computational issues see Grant, Liu, and Stoeckert (2005).
4 Estimation procedures

Given a pre-specified type I error \( \alpha \) the choice of algorithm can be crucial since we denote those genes 'significant' that have an estimated \( FDR \) level below \( \alpha \). In this paper we demonstrate how the identification of differentially expressed ('significant') genes depends on the estimation procedure, on computational issues, and on the default parameter settings. We restrict ourselves to permutation-based procedures where the resampling step is used for estimation of the joint distribution of the test statistic. This means to permute entire columns of the expression matrix, making the group variable independent of the expression levels while preserving the correlation structure present in the expression data (Ge, Dudoit, and Speed, 2003).

Two of the \( R \) procedures considered here are implementations of the well-known SAM approach, \texttt{samr} (Chu et al., 2005) and \texttt{siggenes} (Schwender, Krause, and Ickstadt, 2003). The third procedure had been recently proposed by Grant, Liu, and Stoeckert (2005) and was implemented by us in \( R \). A substantial difference between \texttt{samr} and \texttt{sam} of the \texttt{siggenes} package is the type of rule for estimating the proportion of truly null hypotheses which in turn leads to different \( FDR \) estimates. While the \texttt{samr} rule is based on the number of observed test statistics belonging to the interquartile range of all permuted test statistics (Tusher et al., 2001), the \texttt{sam} rule is based on the method of Storey and Tibshirani (2003) where the estimation of the distribution of the \( p \)-values is evaluated via natural cubic splines. Another important difference is the fact that \texttt{siggenes}' \texttt{sam} function allows us to either use the mean or median value of falsely significant genes obtained from the set of \( B \) permutation steps when estimating the \( FDR \). Further it should be mentioned that from a computational point of view the core function of the \texttt{samr} package is less preferable because of its high memory requirements due to the storage of numerous intermediate results. The idea of Grant's procedure is to estimate the \( FDR \) for an appropriate set of values (can be thought of as thresholds) covering the range of observed test statistics and then to pick the value satisfying the pre-specified \( \alpha \) level. For each \( k \) in the given range we estimate the approximation of the \( FDR \) with \( V_k / R_k \) where \( R_k \) is the number of observed test statistics greater than \( k \) and \( V_k \) is the estimate of the number of truly null genes among those \( R_k \) genes in an iterative algorithm. This algorithm refines the initial estimates resulting from the set of \( B \) permutations of the data (for details see Grant, Liu, and Stoeckert, 2005).

All three procedures apply to the (modified) \( t \)-statistic for which the so-called 'fudge factor' \( s_0 \), a tuning constant, needs to be specified. In both \( R \) implementations of SAM, \( s_0 \) can be either estimated via the popular algorithm of Tusher, Tibshirani, and Chu (2001) or specified by the user. For the third \( FDR \) controlling mechanism it is suggested to use a meaningful set of \( s_0 \) values and then to select one appropriate value (a subjective user
choice). The influence of $s_0$ on the results is of additional interest in our paper.

5 The simulation study

In our simulation study we analyzed 11 artificial data sets, one having independent rows (i.e. genes) and 10 others derived from the independent one with the intention to mimic co-regulation (correlation) of selected genes. All the data sets consist of 3000 genes and 20 samples (10 in each group) with 2700 genes thought of as truly null and hence distributed $N(0,1)$. 300 of the simulated genes are assumed to be differentially expressed in the second group, 100 up-regulated and 200 down-regulated. The up-regulated genes of the independent data set were constructed as follows: the first 25 being distributed according to $N(1,1)$, the next 25 genes following $N(1,2)$, the subsequent 25 genes being derived from $N(2,1)$, and the last 25 from $N(2,2)$. The design for down-regulated genes is analogue: the first 50 being distributed according to $N(-1,1)$, the next 50 being derived from $N(-1,2)$, the subsequent 50 from $N(-2,1)$, and the last 50 from $N(-2,2)$. The cor-
related data sets were derived from the independent one in the following manner: from each of the differentially expressed groups of genes \(N(1,1), N(1,2), \ldots\) five genes were picked at random and then duplicated to form a new group with known correlation structure. For up-regulated features each gene was copied five times (to have a total number of 25 genes), whereas for down-regulated features each gene was copied 10 times (to complete the total number of 50 genes). Afterwards random noise following \(N(0,0.1)\) was added to ensure that the expression measurements of the five (respectively 10) related genes show biological (i.e. stochastic) variation.

In total we compared four permutation-based approaches implemented in R. For all artificial data sets we performed 3,000 permutations unless more were required for the sake of stable results. The first procedure considered is the one due to Grant, Liu and Stoeckert (2005) and is denoted ‘A’ in the plots. The second method studied is \texttt{siggenes} applying the mean, and the third is \texttt{siggenes} applying the median, called ‘B’ respectively ‘C’ in the plots. Finally we compared these results with those from the well-known SAM package called \texttt{samr}, denoted ‘D’.

We studied the power (specific ability of selecting genes as significant given a pre-specified FDR significance level \(\alpha\)) and the bias of these procedures. For all data sets the following values of the tuning constant \(s_0\) were adopted: 0, 0.5, 1, and 5, and in addition those \(\hat{s}_0\) provided by \texttt{samr} and \texttt{siggenes} (see below). The chosen FDR significance levels are 0.05 and 0.10. The computations were performed on a MS-Windows Pentium platform with 1,024 MB RAM.

6 Results and conclusions

Before we discuss the results we wish to point out an unexpected discrepancy of the \(\hat{s}_0\) estimates between \texttt{samr} and \texttt{siggenes}. Despite the fact that their authors refer to the same (original) algorithm due to Tusher et al. (2001), we obtained substantially different estimates for the tuning constant for six out of 11 analyzed data sets.

Let us now study the multiple box plots in FIGURES 1 to 4. FIGURES 1 and 3 summarize the results for the achieved FDR estimates in comparison with the nominal \(\alpha\) level of 0.05 and 0.10, respectively. FIGURES 2 and 4 give the numbers of differentially expressed (‘significant’) genes (the counts for up- and down-regulated genes are collapsed) obtained for the nominal \(\alpha\) level of 0.05 and 0.10, respectively. In each figure the results for the correlated data (box plots) are combined with those for the independent data (bullets). There is a general tendency that the obtained numbers of differentially expressed genes for the independent data are located in the upper quartile of those for correlated data. One can conclude that all procedures can cope with correlated data, although with a slight tendency to call less genes.
FIGURE 2. Number of differentially expressed genes obtained from the four procedures (for the description of 'A', 'B', 'C', and 'D' see text) for selected $s_0$ under nominal $\alpha = 0.05$ (independent data: •; correlated data: box plots).

FIGURES 2 and 4 show us that the samr function and the sam function applying the median behave similarly and are outperforming the other two procedures with respect to power, particularly for larger values of $s_0$. Apart from this comparative finding, the usefulness of Grant’s procedure is questionable because we had to perform the enormous amount of 10,000 permutations to obtain stable results. As far as bias is concerned (FIGURES 1 and 3) there is an overall tendency towards higher accuracy of all procedures under $\alpha = 0.10$. For $s_0 = 5$ median-based procedures are less biased compared to mean-based methods. However the latter are quite precise for small values of the tuning parameter and can be substantially biased in downward direction for larger values.

When considering the impact of the tuning parameter on the $t$-statistic we would like to point out the following fact: only in two cases (out of 11) the number of significant genes was maximized for one of the $s_0$ obtained from the algorithm of Tusher et al. (2001), otherwise the maximum number of significant genes was achieved for values equal to one or five. Thus the default values obtained from samr and siggenes not only differ but can be far from optimal with respect to maximum power. From our results (tables not displayed here) one can also conclude that there is an interaction...
FIGURE 3. FDR estimates obtained from the four procedures (for the description of 'A', 'B', 'C', and 'D' see text) for selected $s_0$ under nominal $\alpha = 0.10$ (independent data: *; correlated data: box plots).

between the magnitude of $s_0$ and the distributional features of the truly null expressed genes, i.e. the change in the spectrum of called genes with increasing $s_0$. Theoretically speaking, for small tuning constants a large variance of some genes can mask their absolute difference in the means while a tiny variance of other genes can cause significance. This may be the primary reason for improved results for larger values of $s_0$. In summary we can say that the original SAM procedure implemented in samr performs well for both independent and correlated data, however sam of the siggenes package with the median as the estimate of falsely called genes can outperform it. The choice of the tuning constant is quite critical and default values provided by the SAM procedures are often far from optimal.

References


FIGURE 4. Number of differentially expressed genes obtained from the four procedures (for the description of ‘A’, ‘B’, ‘C’, and ‘D’ see text) for selected $s_0$ under nominal $\alpha = 0.10$ (independent data: ●; correlated data: box plots).


Analyzing Irish suicide rates with mixture models

Nick Sofroniou\textsuperscript{1}, Jochen Einbeck\textsuperscript{2} and John Hinde\textsuperscript{2}

\textsuperscript{1} Educational Research Centre, St Patrick’s College, Dublin, Ireland
\textsuperscript{2} Department of Mathematics, National University of Ireland, Galway, Ireland

Abstract: In the analysis of morbidity and mortality data, variance component models are commonly used to provide an improvement in the estimation of rates for small regions which typically show large variability. This article investigates Irish suicide data using Poisson mixed models. The random effect distributions are estimated using Nonparametric Maximum likelihood which allows the calculation of shrinkage estimates from the posterior probability estimates of the EM algorithm, as well as the construction of ‘league tables’. As these models are inefficient in the case of spatial dependency, we investigate the addition of spatial autocorrelation terms based on neighboring average crude rates and standardized mortality ratios, as well as gender-specific versions of these. We consider models for the average crude rate as well as for the relative risk. A close correspondence between fitted values from both types of models suggests that information concerning within-region variability, incorporated in the parameters of the average crude rate model, appears indirectly in the simpler relative risk model by means of the expected values used in the offset term of the latter.

Keywords: Generalized linear models; random effects; nonparametric maximum likelihood; spatial autocorrelation; suicide rates.

1 Introduction

The use of generalized linear models with random effects is already well established in the analysis of morbidity and mortality data. Administrative regions defined in geographical terms have reported counts of cause of death or illness and the aim is to model the variation in these. Calculating separate estimates of risk for each area may result in small regions having estimates with large variability, leading to small-area estimation problems (e.g., Longford, 2005). Variance component models enable the generation of empirical Bayes shrinkage estimates to improve the estimation of local risk (Aitkin, 1996b).

Assume we have a division of some region into \( m \) districts with population sizes \( n_i, i = 1, \ldots, m \) and counts \( Y_i, i = 1, \ldots, m \), and a further division into subpopulations \( j = 1, \ldots, J \) (e.g., certain gender/age groups) with explanatory vectors \( x_{ij} \), observed counts \( Y_{ij} \), and sizes \( n_{ij} \), such that
\( \sum_j Y_{ij} = Y_i \) and \( \sum_j n_{ij} = n_i \). The observed mortality/morbidity counts \( Y_{ij} \) are commonly assumed to follow a Poisson distribution with mean \( \mu_{ij} \), which can be either specified using the rate \( \lambda_{ij} \), giving \( \mu_{ij} = n_{ij} \lambda_{ij} \), or the relative risk \( \theta_{ij} \), implying that \( \mu_{ij} = E_{ij} \theta_{ij} \), with \( E_{ij} \) being the expected number of cases obtained from some reference population (Ahlbom, 1993). Alternatively, the models can be based on a binomial distribution as in Aitkin (1996b). The Poisson distribution is the more natural choice if the occurrence of the death/disease is a rather rare event.

If the number \( m \) of districts is quite high (say, more than six or seven), modelling the regional heterogeneity as a fixed effect would require a quite large number \( (m - 1) \) of additional model parameters. This can be avoided by using random effects \( z_i, i = 1, \ldots, m \). For the two cases mentioned above, we use a log-linear model for the parameter of interest,

\[
\begin{align*}
\{ \log(\lambda_{ij}) \} &= \beta' x_{ij} + z_i, \\
\log(\theta_{ij}) &= \beta' x_{ij} + z_i,
\end{align*}
\]

yielding the generalized random effect models

\[
\log(\mu_{ij}) = \text{offset} + \beta' x_{ij} + z_i,
\]

with offsets \( \log(n_{ij}) \) or \( \log(E_{ij}) \), respectively. One observes from (2) that both families of models actually only differ by the offset, and hence can be represented within a larger family where the two offsets are present in the linear predictor, each multiplied by an indicator variable to select the relevant offset. Thus, although the autocorrelation terms and offsets being compared may differ, the likelihoods will still be on the same scale and they can be compared with each other using their disparities (i.e., \(-2 \log L, \) with \( L \) being the likelihood).

In this article we analyze Irish suicide data using both models specified in (1). The rates \( \lambda_{ij} \) correspond in our setting to death rates due to “suicide or intentional self-harm”. The tool used for analysis of the data is the Nonparametric Maximum Likelihood (Aitkin, 1996a). The paper can be seen on the one hand as a methodological addition to the findings in Aitkin (1996b), as we extend the idea of empirical Bayes shrinkage to situations where one or more covariates are present, and on the other hand as a complement to the models introduced in Biggeri et al. (2000), as we explore the ample ground between modelling a random spatial autocorrelation term and not modelling spatial autocorrelation at all.

## 2 Irish suicide data

The data considered here describe the mortality due to suicide and intentional self-harm in the Republic of Ireland from 1989–1998, obtained from the All Ireland Mortality Database (Institute of Public Health in Ireland, etc.).
FIGURE 1. Map of Health Boards and Cities for the Republic of Ireland. The excluded regions of Northern Ireland are shown in dark grey. The ‘−’ sign indicates that a city is excluded from its health board.

2005). This database divides the Republic of Ireland into 13 ‘health regions’ (the 8 former health boards which existed during this period, and the cities Cork, Dublin, Galway, Limerick, and Waterford extracted from these health boards; see Fig. 1). The data are graphically displayed in Fig. 2 (left) and are part of the R package npmlreg (Einbeck et al., 2006). We will use the explanatory variables gender, age, a suitable measure of regional autocorrelation, and a cluster-level random effect to account for the regional heterogeneity (e.g., arising from regions with big/small populations, outliers etc.). This leads to a two-level model, also called a variance component model, where the clustering variable is the health region ID. The age variable is a factor with four categories from 0–29 (reference category), 30–39, 40–59, and 60+ years.

For each region \(i = 1, \ldots, 13\) and each subpopulation \(j = 1, \ldots, 8\) (defined by a certain gender/age combination), we have a total count of suicides \(Y_{ij}\) over the 10 years. Further, the subpopulation sizes \(n_{ij}\) are available, as well as the standardized mortality ratios (SMR), i.e., the ratio observed/expected number of deaths, from which the \(E_{ij}\) are immediately obtained.
3 Modelling suicide rates

We firstly focus on the model for the rate $\lambda$. The ‘core’ model

$$\log(\lambda_{ij}) = \alpha + \beta_1 \cdot \text{sex}_{ij} + \beta_2 \cdot \text{age}_{2,ij} + \beta_3 \cdot \text{age}_{3,ij} + \beta_4 \cdot \text{age}_{4,ij} \equiv \alpha + \eta_{ij} \quad (3)$$

gives a disparity of $-2 \log L = 793.8$, with all five estimated parameters being highly significant. Next, we replace the constant intercept $\alpha$ by a regional random effect $z_i$, assuming that all individuals living within one health region share a common intercept. The NPML approach approximates the unknown and unspecified distribution of the random effects by a discrete mixture, yielding mass points $z_1, \ldots, z_k$ and masses $\pi_1, \ldots, \pi_k$.

Fitting a model with $k = 3$ mass points, the disparity already drops to $697.2$, and does not fall significantly when increasing $k$ further.

To improve this result, we construct average ‘neighboring crude rates’ $r_i = \sum_{\ell \in S_i} Y_{\ell}/\sum_{\ell \in S_i} n_{\ell}$, where $S_i$ is the set of regions adjacent to the $i-$th region. Including this variable as a fixed effect, the model can be formulated as

$$\log(\lambda_{ij}) = \eta_{ij} + \gamma \cdot r_i + z_i,$$

and the disparity drops to $691.3$. Using a random instead of a fixed coefficient for the autocorrelation term, one achieves a slight additional decrease in disparity down to $689.8$. However, using gender-specific neighboring rates $r_{ij}$ and a fixed parameter $\gamma$, nearly the same reduction can be achieved, yielding the disparity $690.1$. Combining these ideas using a random gender-specific autocorrelation term one gets a further slight improvement towards $688.8$ (all models using $k = 3$). The addition of interaction terms between age and sex as fixed effects yields a further improvement in the model, with the deviance dropping to $646.8$ for fixed $\gamma$, and to $645.1$ for a random coefficient $\gamma$. Summarizing, largely independent of the order of the inclusion of the terms, we get a disparity reduction of about 95 points for the regional random effect, of about 45 points for the interaction, and of about 10 points for the regional autocorrelation.

One nice feature of NPML estimation is that the posterior probability that unit $i$ stems from cluster $k$ corresponds to the weights in the final iteration of the EM algorithm. Firstly, this enables us to calculate empirical Bayes predictions from posterior estimates of the random effects combined with the fixed part of the linear predictor (Aitkin, 1996b, did this for the case of posterior means from a binomial model without covariates), shrinking the – highly variable – crude observations by ‘borrowing’ information from similar regions. The results are shown in Fig. 2 (right) for the fixed gender-specific autocorrelation model with age.gender interaction terms, which we consider as our favorite model from the average crude rate family of models, for this data. One observes that particularly the rates for the cities – based on small population sizes compared to the health boards – are considerably shrunk. Secondly, one can classify the regions into clusters according to the
mass point with the highest posterior probability (for use in ‘league table’ type comparisons and performance monitoring). In the case of a random intercept model with covariates we use the term posterior intercept for the posterior estimate of the random effect term. We use a convention of assigning a cluster to a mass point component if \( p \geq 0.90 \) and confidently excluding it from a component if \( p < 0.01 \) (see Table 1). For a model which incorporates relevant explanatory variables, these classifications are informative about clusters with either excess or exemplary rates of mortality, conditional on the fitted model and the observed data. It turns out that only the health region ‘EHB minus Dublin’ is assigned to mass point 1 (as a region with very few suicides), the city Cork and the region ‘SEHB minus Waterford’ are classified to mass point 3 as regions with a large number of suicides, and all other regions are assigned to the intermediate mass point 2, except the cities of Waterford and Limerick of which we are limited to inferences regarding the mass points that they can be excluded from (Limerick is excluded from the low suicide rate and Waterford from the high suicide rate).

4 Modelling relative risks

A similar analysis as in Section 3 is conducted using the relative risk parameter \( \theta \) as model parameter. We make use of the simpler ‘core’ model \( \log(\theta_{ij}) = \alpha \), giving the disparity 754.4. One might wonder at this point why we get a better disparity compared to model (3), using a model without covariates. The reason for this is that the information about the explana-
TABLE 1. Posterior probabilities for suicide data modelled with crude rate (left) and relative risk (right) as model parameters.

<table>
<thead>
<tr>
<th></th>
<th>Average crude rate</th>
<th>Relative risk</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intercept</td>
<td>Masspoints</td>
<td>Intercept</td>
</tr>
<tr>
<td>Intercept</td>
<td>−8.910</td>
<td>0.118</td>
<td>−0.576</td>
</tr>
<tr>
<td>SEHB − Waterford</td>
<td>−8.31</td>
<td>1.00</td>
<td>0.03</td>
</tr>
<tr>
<td>Cork City</td>
<td>−8.31</td>
<td>1.00</td>
<td>0.02</td>
</tr>
<tr>
<td>Limerick City</td>
<td>−8.46</td>
<td>0.68</td>
<td>0.32</td>
</tr>
<tr>
<td>NEHB</td>
<td>−8.53</td>
<td>1.00</td>
<td>−0.19</td>
</tr>
<tr>
<td>Dublin City</td>
<td>−8.53</td>
<td>1.00</td>
<td>−0.19</td>
</tr>
<tr>
<td>SHB − Cork</td>
<td>−8.53</td>
<td>1.00</td>
<td>−0.19</td>
</tr>
<tr>
<td>Mid WHB − Limerick</td>
<td>−8.53</td>
<td>1.00</td>
<td>−0.19</td>
</tr>
<tr>
<td>Midland HB</td>
<td>−8.53</td>
<td>1.00</td>
<td>−0.19</td>
</tr>
<tr>
<td>NWBH</td>
<td>−8.53</td>
<td>1.00</td>
<td>−0.19</td>
</tr>
<tr>
<td>WHB − Galway</td>
<td>−8.53</td>
<td>1.00</td>
<td>−0.19</td>
</tr>
<tr>
<td>Galway City</td>
<td>−8.56</td>
<td>0.07</td>
<td>0.92</td>
</tr>
<tr>
<td>Waterford City</td>
<td>−8.71</td>
<td>0.47</td>
<td>0.53</td>
</tr>
<tr>
<td>EHB − Dublin</td>
<td>−8.91</td>
<td>1.00</td>
<td>−0.58</td>
</tr>
</tbody>
</table>

Posterior probabilities: $p \geq 0.95$, $0.90 \leq p < 0.95$, $p < 0.90$.

...tory variables – including the interaction – is essentially contained in the expected values $E_{ij}$, which goes into the model as an offset according to equation (2). Hence, the 40 points improvement compared to model (3) stems from the indirect inclusion of main effects and an interaction term. Carrying out an analysis along the same lines as in Section 3, our favorite model,

$$\log(\theta_{ij}) = \gamma \cdot r_{ij} + z_i,$$

again turns out to contain a random intercept $z_i$ for regions and a fixed gender-specific autocorrelation term. Note that the $r_{ij}$ are now computed as average neighboring SMRs. The disparity 647.5 of this model is only very slightly worse than in Section 3 (646.8), given that we save 7 degrees of freedom, just by employing another offset!

The fitted values are very similar to those of the final average crude rate model and show considerable shrinkage for the city regions. The strong agreement between the fitted values of the models (Fig. 3), despite the omission of the age and sex variables for the relative risk model, supports our statement above that information concerning variation in rates within a region is incorporated in the expected value offset used with the latter family of models. Classifying the regions into mass point components based on their posterior probabilities (shown in Table 1) again indicates that Cork City and ‘SEHB minus Waterford’ are assigned to the high suicide rate mass point 1, and ‘EHB minus Dublin’ is the only region assigned to the low suicide rate mass point 3. The other regions are assigned to the intermediate rate mass point 2, apart from Waterford City which is excluded from the high rate and Limerick City which is excluded from the low rate, but both of which have posterior probabilities spread across two mass points.
5 Conclusion

In the context of the Irish suicide mortality dataset, a spatial autocorrelation appears to be separately identifiable, in addition to the random-effect for regional heterogeneity. Summarizing the present findings, we conclude i) Modelling regional heterogeneity with spatial random effects improves the model fits greatly. ii) Further improvements can be gained including a spatial autocorrelation term. iii) Unlike Biggeri et al. (2000), we do not observe much gain in using a random coefficient for the autocorrelation term. iv) The relative-risk models incorporate information about variation within regions through the expected values, rather than through the additional covariate terms. The average crude rate modelling approach may be preferable when there is interest in evaluating the effects of explanatory variables, e.g., to inform our understanding of the data generating process. Further the average crude rate approach allows continuous covariates and finer groupings into factors, since the problem of counts with corresponding SMR values of zero does not arise in that case.

Due to differences in administration and health policy, the bordering regions of Northern Ireland were omitted in the calculation of the autocorrelation terms for the adjacent regions in the Republic of Ireland. Further development along the lines of the present analysis might incorporate the 6 regions of the North and examine whether autocorrelations which include cross-border effects improve the fit of the model. National differences in the rates across the two sets of counties could be allowed for by means of additional interaction terms, though the number of regions in the North is small.

We finish with a word of caution: though we did not observe computational problems in fitting these models neither using a GLIM nor using an R implementation of NPML, certain problems can arise in jointly modelling both
heterogeneity and spatial dependence between regions, as noted by Aitkin (1999), e.g., in some cases a joint distribution of spatial random effects for each region may be singular given a very high intra-area correlation.

Acknowledgments: This work was partly supported by Science Foundation Ireland Basic Research Grant 04/BR/ M0051.

References


Statistical analysis of selected European stock indices

Jiří Trešl and Dagmar Blatná

1 University of Economics Prague, W. Churchill Sq. 4, Prague, Czech Republic, tresl@vse.cz, blatna@vse.cz

Abstract: The analysis of both Central-European and West-European stock market indices was performed. Time series of logarithmic returns were always characterized by higher kurtosis and non-normal fitted distribution. Second, ACF values are rather small, but still persistent up to relatively high order. Finally, due to obvious heteroscedasticity, mixed AR-GARCH models are suitable and GARCH(1,1) model seems to be quite satisfactory.

Keywords: Stock returns; financial time series; GARCH models.

1 Introduction

The main aim of this study is to compare the behaviour of stock markets both in West Europe and in Central-European transitionary markets. Input data are given as daily values of corresponding stock indices $x_t$ during the period 1998-2004, i.e. 1655 daily values. As for indices, we selected UKX (GB), DAX (Germany), CAC (France) and MIBTEL (Italy) and, on the other hand, PX-50 (Czech Republic), SKSM (Slovak Republic), WIG (Poland) and BUX (Hungary). The subject of our analysis were logarithmic returns (further only returns) expressed as percentage and computed as

$$y_t = 100(\log x_t - \log x_{t-1})$$

This is a common practice in financial time series analysis, because continuously compounded multiperiod return is simply the sum of continuously compounded one-period returns involved [1].

2 Descriptive methods

First, some elementary summary statistics related to daily returns were computed. The results obtained are compiled in Table 1. Clearly, there are no distinct differences in unconditional variability, i.e. standard deviation and mean absolute deviation. Further, extreme values for BUX and WIG are roughly twice the size of ones for PX-50 and UKX. Notice that the kurtosis is always positive, i.e. the empirical distribution
TABLE 1. Summary statistics.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>R_BUX</th>
<th>R_CAC</th>
<th>R_DAX</th>
<th>R_MIBTEL</th>
<th>R_PX-50</th>
<th>R_SKSM</th>
<th>R_UKX</th>
<th>R_WIG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>0.05</td>
<td>0.04</td>
<td>0.06</td>
<td>0.04</td>
<td>0.03</td>
<td>0.00</td>
<td>0.02</td>
<td>0.04</td>
</tr>
<tr>
<td>Stddev</td>
<td>1.86</td>
<td>1.64</td>
<td>1.85</td>
<td>1.45</td>
<td>1.38</td>
<td>1.47</td>
<td>1.30</td>
<td>1.73</td>
</tr>
<tr>
<td>MAD</td>
<td>0.89</td>
<td>0.91</td>
<td>1.03</td>
<td>0.73</td>
<td>0.78</td>
<td>0.61</td>
<td>0.70</td>
<td>0.84</td>
</tr>
<tr>
<td>Minimum</td>
<td>-15.53</td>
<td>-7.02</td>
<td>-8.73</td>
<td>-10.47</td>
<td>-7.08</td>
<td>-11.48</td>
<td>-5.59</td>
<td>-13.54</td>
</tr>
<tr>
<td>Maximum</td>
<td>16.32</td>
<td>13.97</td>
<td>15.27</td>
<td>14.43</td>
<td>6.22</td>
<td>5.96</td>
<td>6.82</td>
<td>13.53</td>
</tr>
<tr>
<td>Range</td>
<td>31.85</td>
<td>20.99</td>
<td>24.00</td>
<td>24.90</td>
<td>13.30</td>
<td>17.44</td>
<td>12.41</td>
<td>27.07</td>
</tr>
<tr>
<td>L quartile</td>
<td>-0.86</td>
<td>-0.89</td>
<td>-1.04</td>
<td>-0.68</td>
<td>-0.73</td>
<td>-0.54</td>
<td>-0.70</td>
<td>-0.85</td>
</tr>
<tr>
<td>U quartile</td>
<td>0.93</td>
<td>0.91</td>
<td>1.05</td>
<td>0.77</td>
<td>0.83</td>
<td>0.65</td>
<td>0.72</td>
<td>0.86</td>
</tr>
<tr>
<td>IQ range</td>
<td>1.79</td>
<td>1.80</td>
<td>2.09</td>
<td>1.45</td>
<td>1.56</td>
<td>1.19</td>
<td>1.42</td>
<td>1.71</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.29</td>
<td>0.30</td>
<td>0.20</td>
<td>0.18</td>
<td>-0.04</td>
<td>-0.67</td>
<td>-0.01</td>
<td>0.36</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>12.87</td>
<td>4.61</td>
<td>4.08</td>
<td>9.17</td>
<td>1.90</td>
<td>6.30</td>
<td>2.01</td>
<td>9.99</td>
</tr>
</tbody>
</table>

exhibits greater peakedness in comparison with normal one. This tendency to more pronounced peaks and heavy tails seems to be a persistent feature in financial time series [1].

As for a suitable theoretical model for returns distribution fitting, the following types were examined: Laplace, logistic, normal and Student’s t distributions [2]. The results obtained are summarised in Table 2.

TABLE 2. P-values for distribution fitting using chi-square test.

<table>
<thead>
<tr>
<th></th>
<th>R_BUX</th>
<th>R_CAC</th>
<th>R_DAX</th>
<th>R_MIBTEL</th>
<th>R_PX-50</th>
<th>R_SKSM</th>
<th>R_UKX</th>
<th>R_WIG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laplace</td>
<td>0.079</td>
<td>0.000</td>
<td>0.000</td>
<td>0.012</td>
<td>0.005</td>
<td>0.000</td>
<td>0.007</td>
<td>0.302</td>
</tr>
<tr>
<td>Logistic</td>
<td>0.446</td>
<td>0.014</td>
<td>0.007</td>
<td>0.000</td>
<td>0.175</td>
<td>0.000</td>
<td>0.170</td>
<td>0.002</td>
</tr>
<tr>
<td>Normal</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Student</td>
<td>0.356</td>
<td>0.000</td>
<td>0.000</td>
<td>0.027</td>
<td>0.280</td>
<td>0.000</td>
<td>0.433</td>
<td>0.014</td>
</tr>
</tbody>
</table>

Thus, logistic and Student’s t distributions proved to be the best ones, whereas normal distribution failed in all cases, because it has systematically lower kurtosis.

To reveal possible relations among individual returns, partial correlations were computed with the results given in Table 3.

3 Modelling

The next step is the possibility of modelling of return time series. Though are ACF values small, they seem to be persistent up to higher orders. Further, they tend to have the same sign at certain order (e.g. the
TABLE 3. Partial correlations. (All values are statistically significant at 5% level.)

<table>
<thead>
<tr>
<th></th>
<th>R_BUX</th>
<th>R_CAC</th>
<th>R_DAX</th>
<th>R_MIBTEL</th>
<th>R_PX-50</th>
<th>R_KSM</th>
<th>R_UKX</th>
<th>R_WIG</th>
</tr>
</thead>
<tbody>
<tr>
<td>R_BUX</td>
<td></td>
<td>0.092</td>
<td>0.135</td>
<td>0.285</td>
<td>-0.016</td>
<td>0.062</td>
<td>0.259</td>
<td></td>
</tr>
<tr>
<td>R_CAC</td>
<td></td>
<td></td>
<td>0.373</td>
<td>0.067</td>
<td>-0.041</td>
<td>0.494</td>
<td>-0.084</td>
<td></td>
</tr>
<tr>
<td>R_DAX</td>
<td></td>
<td></td>
<td></td>
<td>0.042</td>
<td>0.009</td>
<td>0.092</td>
<td>0.041</td>
<td></td>
</tr>
<tr>
<td>R_MIBTEL</td>
<td></td>
<td></td>
<td></td>
<td>0.023</td>
<td>0.017</td>
<td>0.005</td>
<td>0.114</td>
<td>0.014</td>
</tr>
<tr>
<td>R_PX-50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.020</td>
<td>0.032</td>
<td>0.176</td>
<td></td>
</tr>
<tr>
<td>R_KSM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.054</td>
<td>0.061</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R_UKX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.033</td>
<td></td>
</tr>
<tr>
<td>R_WIG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

On the other hand, ACF computed for absolute values of returns reaches typical values 0.15 and decreases very slowly, so that all values up to lag = 20 are statistically significant at 5% level. This is a crucial difference in comparison with ACF of white noise process. For this reason, we used combined AR-GARCH models suitable for the modelling in the presence of heteroscedasticity [1], [3]. The governing equations are

\[
\begin{align*}
y_t &= \varphi_1 y_{t-1} + \ldots + \varphi_m y_{t-m} + \epsilon_t \\
\epsilon_t &= \sigma_t \epsilon_t \\
\sigma_t^2 &= \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \\
e_t \text{ is } N(0,1)
\end{align*}
\]

where \(\sigma_t\) denotes conditional standard deviation and \(\epsilon_t\) is normal white noise. In our case, the simplest GARCH(1,1) model was quite efficient and in most cases, only one autoregressive term was statistically significant. The results obtained are presented in Table 4.

Clearly, there is strong direct dependence of conditional standard deviation on its previous value, manifested itself by large values of \(\beta\) parameter. Second, \(\beta\) values vary only slightly among individual stocks returns.

4 Conclusions

As for returns summary statistics, the kurtosis of empirical distribution was always positive and, therefore, logistic or Student’s t distributions proved to be always better fitted in comparison with normal one; the half of values occurs roughly within 1%. Further, unconditional standard deviation
TABLE 4. Parameters of AR-GARCH models used.

<table>
<thead>
<tr>
<th></th>
<th>$\phi(01)$</th>
<th>$\phi(02)$</th>
<th>$\phi(04)$</th>
<th>$\phi(08)$</th>
<th>$\phi(10)$</th>
<th>$\omega$</th>
<th>$\alpha$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>R$_{BUX}$</td>
<td>-0.071</td>
<td>0.055</td>
<td>-0.075</td>
<td>-0.084</td>
<td>0.063</td>
<td>0.053</td>
<td>0.926</td>
<td>0.935</td>
</tr>
<tr>
<td>R$_{CAC}$</td>
<td>-0.090</td>
<td>-0.054</td>
<td>0.075</td>
<td>0.084</td>
<td>0.025</td>
<td>0.056</td>
<td>0.911</td>
<td>0.914</td>
</tr>
<tr>
<td>R$_{DAX}$</td>
<td>0.069</td>
<td>0.055</td>
<td>0.075</td>
<td>0.084</td>
<td>0.048</td>
<td>0.076</td>
<td>0.925</td>
<td>0.925</td>
</tr>
<tr>
<td>R$_{MIBTEL}$</td>
<td>-0.057</td>
<td>-0.074</td>
<td>0.106</td>
<td>0.157</td>
<td>0.106</td>
<td>0.072</td>
<td>0.853</td>
<td>0.872</td>
</tr>
<tr>
<td>R$_{PX-50}$</td>
<td>+0.058</td>
<td>+0.045</td>
<td>0.157</td>
<td>0.028</td>
<td>0.157</td>
<td>0.091</td>
<td>0.853</td>
<td>0.890</td>
</tr>
<tr>
<td>R$_{SKSM}$</td>
<td>0.106</td>
<td>0.045</td>
<td>0.028</td>
<td>0.017</td>
<td>0.028</td>
<td>0.056</td>
<td>0.853</td>
<td>0.890</td>
</tr>
<tr>
<td>R$_{UKX}$</td>
<td>-0.074</td>
<td>-0.045</td>
<td>0.017</td>
<td>0.003</td>
<td>0.017</td>
<td>0.093</td>
<td>0.853</td>
<td>0.890</td>
</tr>
<tr>
<td>R$_{WIG}$</td>
<td>0.058</td>
<td>0.045</td>
<td>0.003</td>
<td>0.045</td>
<td>0.003</td>
<td>0.093</td>
<td>0.853</td>
<td>0.890</td>
</tr>
</tbody>
</table>

seems to be very quiet; on the other hand, one can expect sudden onsets of increased volatility during some periods.

The highest partial correlations occur on the one hand among West-European indices CAC, UKX, DAX, MIBTEL; on the other hand, as for Central-European indices, the correlations are weaker and the highest ones occur among BUX, PX-50 and WIG, whereas SKSM returns exhibit some kind of atypical behaviour. The same conclusions were drawn with the use of cluster analysis. The structure of autocorrelation function (ACF) exhibits several typical features. On the one hand, ACF values are rather small, but, on the other hand, they are persistent up to relatively high order.

Finally, as for modelling, the combination of an autoregressive model and GARCH (1,1) model was sufficient to remove present heteroscedasticity. In all cases, there was strong dependence of conditional variance on its previous value. Thus, the ARMA models are not capable to catch features observed and there is strong need for GARCH models.

Acknowledgments: Financial support from GACR (project 402/05/0128) is strongly appreciated. Special thanks are due to Ondřej Krakovič who rendered this text into LaTeX version.

References


An alternative approach to regularization and variable selection in high dimensional regression modelling

Gerhard Tutz\textsuperscript{1} and Jan Ulbricht\textsuperscript{1}

\textsuperscript{1} Institut für Statistik, Ludwig-Maximilians-Universität, Ludwigstr. 33, D-80539 München, Germany

Abstract: For high dimensional regression settings an alternative estimation procedure is proposed which is able to handle highly correlated data. Like the elastic net it encourages a grouping effect, where strongly correlated predictors tend to have similar estimated effects. This effect is obtained by use of a penalty that contains the marginal correlation between predictors. A modified version of the estimator performs variable selection by use of boosting techniques. Unlike usual componentwise boosting methods, refitting is done for pairs of coefficients. It is demonstrated that the methods are strong competitors to common regularization methods like ridge, lasso and elastic net.

Keywords: Lasso; elastic net; ridge estimator; correlation based boosting.

1 Introduction

High dimensional data which are collected in bioinformatics and medicine have typically a very large number of predictors on just some hundred observations. A typical microarray data set has thousands of variables and fewer than 100 samples. This rises new challenges to find appropriate regression models that allow for accurate predictions and still have simple interpretation. We will focus on the linear regression model

\[ y = X\beta + \epsilon \]

where \( y \) is an \( n \)-vector of random responses, \( X \) is an \( n \times p \) design matrix, \( \beta \) a \( p \)-vector of parameters, and \( \epsilon \) an \( n \)-vector of iid random errors. It is well known that ordinary least squares estimates fail in high dimensions, the variance of the estimates becomes large and mean squared error (MSE) is inflated yielding poor performance. Starting with ridge regression (Hoerl & Kennard, 1970) various shrinkage methods have been proposed to achieve better prediction. A generic form of penalized estimates is given by the criterion

\[ PLS(\beta) = |y - X\beta|^2 + P(\beta) \]
where $P(\beta)$ is a penalty term that penalizes the parameter vector. The bridge penalty model proposed by Frank & Friedman (1993) uses

$$P(\beta) = \lambda \sum_{i=1}^{n} |\beta_j|^\gamma, \quad \gamma > 0$$

where $\lambda \geq 0$ is a tuning parameter. Special cases are ridge regression which uses $\gamma = 2$ and the lasso which uses $\gamma = 1$ (Tibshirani, 1996). A distinct advantage of the lasso is that in contrast to ridge regression (and the bridge penalty model with $\gamma > 0$) it does variable selection. While ridge regression shrinks all the parameter components towards zero, the $L_1$-penalty of the lasso tends to select variables yielding a parsimonious representation of the model.

More recently, Zou & Hastie (2005) pointed out several disadvantages of the lasso. In particular the lasso selects at most $n$ variables before it saturates. If correlation between predictors is high it has been observed that the prediction performance is dominated by ridge regression and if there is a group of variables among which correlations are very high, then the lasso tends to select just one of the group.

A prediction method that seems to have better performance is the elastic net proposed by Zou & Hastie (2005). The penalty term is a compromise between ridge and lasso and has the form

$$P(\beta) = \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p} |\beta_i|^2$$

using two tuning parameters $\lambda_1, \lambda_2 \geq 0$. An effect that the elastic net shows is grouping of predictors. A regression method exhibits the grouping effect if the regression coefficients of a group of highly correlated variables tend to be equal (see Zou & Hastie (2005) and references therein to the importance of the 'grouped variables' situation in the 'large $p$, small $n$' problem).

In the present paper alternative methods are proposed to obtain the grouping effect in high dimensional regression models.

2 The correlation based estimator

Since the grouping effect demands that parameters are similar if covariates are highly correlated it is tempting to incorporate the (marginal) correlations into the penalty term. This is explicitly done in the proposed correlation based penalty

$$P_c(\beta) = \lambda \sum_{i=1}^{p-1} \sum_{j>i} \left\{ \frac{(\beta_i - \beta_j)^2}{1 - \rho_{ij}} + \frac{(\beta_i + \beta_j)^2}{1 + \rho_{ij}} \right\}$$

$$= 2\lambda \sum_{i=1}^{p-1} \sum_{j>i} \frac{\beta_i^2 - 2\rho_{ij}\beta_i\beta_j + \beta_j^2}{1 - \rho_{ij}}$$

(1)
where \( \varrho_{ij} \) denotes the correlation between the \( i \)th and the \( j \)th predictor. From the first form it is immediately seen that for strong positive correlation \( (\varrho_{ij} \to 1) \) the term \( (\beta_i - \beta_j)^2/(1 - \varrho_{ij}) \) becomes dominant yielding estimates \( \hat{\beta}_i \approx \hat{\beta}_j \). For strong negative \( (\varrho_{ij} \to -1) \) the second term dominates yielding estimates \( \hat{\beta}_i \approx -\hat{\beta}_j \).

\[ P_c(\beta) \] may be transformed to
\[ P_c = \lambda (\beta^T D W_1 D \beta + \beta^T A^T W_2 A \beta) \] where \( W_1 = diag(1/(1 - \varrho_{12}), 1/(1 - \varrho_{13}), \ldots) \) is a \( (m \times m) \) diagonal matrix, with \( m = n(n-1)/2 \) denoting the number of pairs \( (i,j), i \neq j \), \( W_2 = diag(1/(1 + \varrho_{12}), 1/(1 + \varrho_{13}), \ldots) \), \( D \) specifies the differences and \( A \) is the matrix that specifies the addition of parameters. Therefore with \( W = D^T W_1 D + A^T W_2 A \) one obtains the simple representation
\[ P_c = \lambda \beta^T W \beta. \]

By usual derivation one obtains the solution to the maximization of \( |y - X\beta|^2 + P_c(\beta) \) as
\[ \hat{\beta}_c = (X^T X + \lambda W)^{-1} X^T y. \]

The contour plots of the correlation based estimator and the competitors lasso, ridge and elastic net are given in Figure 1. The constraint region for the ridge penalty is the disk \( \beta_1^2 + \beta_2^2 \leq c \), for the lasso one obtains the diamond \( |\beta_1| + |\beta_2| \leq c \). Since the diamond has distinct corners, if a solution occurs at a corner it means that one parameter \( \beta_j \) is equal to zero. It is seen that contours for ridge and lasso are highly symmetric, \( x_1 = 0 \) is an axis of symmetry as well as \( x_2 = 0 \). In contrast, the constrained region for the correlation based estimator is an ellipsoid which becomes narrower with increasing correlation. In the special case where \( \varrho = 0 \) it is equivalent to ridge regression.

For penalized estimators \( (\lambda > 0) \) and not perfect correlation \( (\varrho_{ij}^2 \neq 1 \text{ for } i \neq j) \) it may be shown that the estimator \( \hat{\beta}_c \) exists and is unique. For perfectly correlated predictors \( (\varrho_{ij}^2 = 1) \) the penalty (1) is not defined. However, it may be shown that the limit \( (\varrho_{ij}^2 \to 1) \) exists. Consequently \( \hat{\beta}_c \) for penalty correlated predictors is defined by the limit.

Zou & Hastie (2005) showed that in the perfect correlation case \( (\varrho_{ij} = 1) \) coefficient estimators are equal if the penalty is strictly convex. This is not the case for the lasso and the lasso does not even have a unique solution. For \( \lambda > 0 \) the correlation based penalty is a strictly convex penalty function and therefore does not suffer from these problems.

The correlation based penalty enforces the grouping effect by explicitly using the correlation between predictions. If one considers variables that are highly correlated as forming a group estimates should be similar and all the variables should be included. In contrast the lasso tends to select one of the group. In simulations it is found that for highly correlated predictors the correlation based predictor often outperforms ridge regression and the lasso.
FIGURE 1. Left panel: Two-dimensional contour plots of the ridge penalty that coincides with the correlation based penalty for \( \varrho = 0 \) (solid line), lasso penalty (dashed line) and elastic net (dotted line). Mid panel: Two-dimensional contour plots of correlation based penalty for three values of positive correlation \( \varrho = 0.5 \) (solid line), \( \varrho = 0.8 \) (dashed line), and \( \varrho = 0.99 \) (dotted line). Right panel: Two-dimensional contour plots of correlation based penalty for three values of negative correlation \( \varrho = -0.5 \) (solid line), \( \varrho = -0.8 \) (dashed line), and \( \varrho = -0.99 \) (dotted line).

3 A boosting version of the penalized correlation estimator

The simple penalized correlation estimator does not perform variable selection. Therefore componentwise boosting is proposed (see Hastie et al. (2001) for an introduction into boosting and Bühlman & Yu (2003)) for the extension to regression settings. Boosting may be seen as iteratively refitting residuals. In componentwise fitting only one component (usually referring to the weight of one predictor) is updated. The procedure suggested here differs in the important aspect that pairs of coefficients are updated. So refitting uses all of the covariates. Let \( X_{ij} \) denote the correspondingly reduced design matrix and \( P_{c,ij} \) denote the correlation based penalty for the two-component vector \( (\beta_i, \beta_j) \). The algorithm is given in the following

Algorithm PairBoost

Step 1: (Initialization)
Set \( \beta^{(0)} = 0, \mu^{(0)} = 0 \).

Step 2: (Iteration)
For \( m = 1, 2, \ldots \)
Alternative approach to regularization

(a) Estimation

Compute the residuals \( r^{(m)} = y - \hat{\mu}^{(m-1)} \) and fit for \( i, j \in \{1, \ldots, p\}, i \neq j \) the model \( r^{(m)} = X_{ij} \hat{b}_{ij} + \epsilon \) by minimizing \( r^{(m)} - X_{ij} \hat{b}_{ij} + P_{c,ij} \). yielding \( \hat{b}_{ij} = (X_{ij}^T X_{ij} + P_{c,ij})^{-1} X_{ij}^T r^{(m)} \)

(b) Selection

Select the pair of variables which has the best fit, yielding

\[
(i_m, j_m) = \arg \min_{i \neq j} |r^{(m)} - X_{ij} \hat{b}_{ij}|^2.
\]

(c) Refit

With \( \hat{b}_{ij} \) given by \( \hat{b}_{ij} = (b_i, b_j) \) the parameter vector is updated by

\[
\hat{\beta}_i^{(m)} = \begin{cases} 
\hat{\beta}_i^{(m-1)} + b_i, & \text{if } i \in \{i_m, j_m\}, \\
\hat{\beta}_i^{(m-1)}, & \text{otherwise},
\end{cases}
\]

yielding the vector \( \hat{\beta}^{(m)} = (\hat{\beta}_1^{(m)}, \ldots, \hat{\beta}_p^{(m)})^T \) and \( \hat{\mu}^{(m)} = \hat{\mu}^{(m-1)} + X_{i_m,j_m} \hat{b}_{i_m,j_m} \).

The algorithm may be stopped by cross-validation methods or by computing the hat matrix and some information as AIC or BIC. The latter may be derived along the lines of Tutz & Binder (2006).

4 Application to prostate data

The prostate data set has been used by several authors to investigate shrinkage estimators (e.g. Tibshirani, 1996, Zou & Hastie, 2005). In the original study Stamey et al. (1989) examined the correlation between the level of prostate specific antigen and eight clinical measures in men who were about to receive a radical prostatectomy. The measures were log cancer volume (lcavol), log prostate weight (lweight), age, log benign prostatic hyperplasia amount (lbph), seminal vesicle invasion (svi), log capsular penetration (lcp), Gleason score (gleason) and percentage Gleason scores 4 or 5 (pgg45). The logarithm of prostate specific antigen (lpsa) has been used as response.

Hastie et al. (2001) divided the data set into a training set with 67 observations and a test set with 30 observations. Based on this split, the correlation based estimator and the PairBoost algorithm are compared with lasso, elastic net, ridge regression and componentwise boosted ridge regression (Tutz & Binder, 2005). It is seen in Figure 2 that the coefficient build ups for ridge regression and the correlation based estimator on rather similar.
The performance of the estimators has been evaluated by considering the median test error for 20 random splits (67 training observations, 30 observations in the test set). The performance was best for the PairBoost (0.554), followed by the componentwise boosted ridge regression (0.569), ridge regression (0.570), the correlation based estimator (0.573), lasso (0.594) and elastic net (0.599), with the median test errors given in parentheses.

5 Simulations

We consider the model $y = X\beta + \epsilon$, where $\epsilon \sim N(0, 3^2)$. The first simulation is based on a similar scenario as example 3 of Zou & Hastie (2005). The simulated data consist of a training set ($n = 100$) an independent validation set (100 observations) and an independent test set (400 observations). The parameter vector is given by

$$\beta^T = (0, \ldots, 0, 2, \ldots, 2, 0, \ldots, 0, 2, \ldots, 2)^T,$$

and the correlation is specified by $\rho(x_i, x_j) = \rho^{|i-j|}$. As predictive measure we consider the mean squared error on the test set.

The parameter vector in the second simulation scenario is given by

$$\beta^T = (3, \ldots, 3, 0, \ldots, 0, 3, \ldots, 3, 0, \ldots, 0, 3, \ldots, 3)^T,$$

and the correlation $\rho(x_i, x_j) = \rho_{ij}$ is given by

$$\rho_{ij} = \begin{cases} 
      1 - 0.01 \cdot |i - j|, & i, j \in \{k, k+1, \ldots, k+4\}, k \in \{1, 6, 11, 16, 21\} \\
      1, & i = j \\
      \epsilon_{ij}, & \text{otherwise,}
   \end{cases}$$

where $\epsilon_{ij}$ are iid $N(0,0.1^2)$. The simulated data consist of a training set ($n = 20$) an independent validation set (20 observations) and an independent test set (80 observations).

Table 1 reports in detail the simulation results. The correlation based estimator performs best for simulation 1 with highly ($\rho = 0.9$) correlated regressors. PairBoost is clearly the best in simulation 2.

6 Concluding remarks

The correlation based penalty has by construction several advantages over the lasso if one wants the grouping effect. In contrast to the elastic net which has a similar focus it has an explicit solution and only one tuning parameter which has to be selected. The disadvantage that variables are not selected may be overcome by using the paired boosting procedure.
TABLE 1. Median mean squared errors on the test sets for the simulations and six methods based on 50 replications. Estimated standard errors in parentheses.

<table>
<thead>
<tr>
<th>Method</th>
<th>Sim. 1, $\rho = 0.9$</th>
<th>Sim. 1, $\rho = 0.7$</th>
<th>Sim. 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ridge penalty</td>
<td>1.04 (0.46)</td>
<td>1.73 (0.47)</td>
<td>3.04 (1.31)</td>
</tr>
<tr>
<td>comp.boost.ridge</td>
<td>1.15 (0.51)</td>
<td>1.77 (0.53)</td>
<td>5.81 (2.96)</td>
</tr>
<tr>
<td>Lasso</td>
<td>1.17 (0.60)</td>
<td>1.42 (0.51)</td>
<td>2.11 (2.36)</td>
</tr>
<tr>
<td>Elastic net</td>
<td>1.10 (0.58)</td>
<td>1.41 (0.50)</td>
<td>2.06 (2.34)</td>
</tr>
<tr>
<td>$p_c$ penalty</td>
<td>0.94 (0.44)</td>
<td>1.64 (0.46)</td>
<td>2.52 (1.25)</td>
</tr>
<tr>
<td>PairBoost</td>
<td>1.07 (0.47)</td>
<td>1.39 (0.53)</td>
<td>1.93 (2.36)</td>
</tr>
</tbody>
</table>

References


FIGURE 2. Coefficient build-ups for prostate data based on six estimation methods: lasso (upper left panel), elastic net (upper right panel) with \( \lambda = 950 \), ridge regression (mid left panel) with varying \( \lambda \), componentwise boosting with ridge penalty (mid right panel) with \( \lambda = 1.84 \) and varying \( m \), penalized least squares (lower left panel) with varying \( \lambda \) and PairBoost (lower right panel) with \( \lambda = 85 \) and varying \( m \).
Restricted predictors for small area totals

Ugarte, M.D.\textsuperscript{1}, Militino, A.F.\textsuperscript{1} and Goicoa, T.\textsuperscript{1}

\textsuperscript{1} Departamento de Estadística e Investigación Operativa, Universidad Pública de Navarra, Campus de Arrosadia, 31006 Pamplona, Spain, lola@unavarra.es

Abstract: The increasing demand for precise estimates in small areas or unplanned domains has promoted the use of small area estimation techniques. One of the most popular methods has been the use of linear mixed models as a basis to obtain empirical best linear unbiased predictors (EBLUP) of population means and totals. However, the estimates provided by statistical offices at some level of aggregation do not necessarily match up with those obtained by adding the estimates derived for small areas using EBLUPs. To obtain the desired concordance, we propose to impose some restrictions in the estimation process of the model parameters. We prove some properties of the new restricted predictor, and obtain the mean squared error using bootstrap procedures. The results are illustrated with real data from the 2002 Basque Country Annual Business Survey.

Keywords: Benchmarking; mixed model; prediction theory; official statistics.

1 Introduction

The increasing demand for precise estimates in local areas has promoted the use of small area techniques in statistical offices. Usually, these agencies must provide estimates for domains that were not originally planned and, as a result, they are poorly represented in the sample or even not represented at all. These domains are called small areas and they usually correspond to small geographic areas, such as counties, municipalities or administrative divisions. Traditionally, the sample sizes are chosen to provide reliable estimates for large regions or aggregates of small areas, but the statistical methods used for large domains can rarely be applied to small ones because the classical direct estimators yield very large standard errors or even they can not be calculated because of the lack of sample in some unplanned domains. Then, the model-based approach to sampling is an alternative over the design-based approach. The main advantage of the model-based approach is that it allows to obtain estimates in small areas with very small sample sizes or even without any sampled observation, by “borrowing information” from related areas, past occasions of both. For a review of small area models see, for example, Rao (2003). The literature in small area estimation is very extensive, and one of the most popular methods is the use of linear mixed models to derive empirical best linear
unbiased predictors (EBLUP). Unfortunately, there are some drawbacks of this methodology making difficult its implementation by statistical agencies. One of them is that the estimates provided by statistical offices at some level of aggregation do not match up with those obtained by adding the estimates derived for small areas with the specific methodology, and then, different estimates for the same quantity are provided. As a consequence, official statistical agencies must adjust the small area estimates to make them coherent with more accurate values at some level of aggregation. This adjustment process is called benchmarking and the more accurate values are called benchmarks. The EBLUP suffers from the deficiency aforementioned. This leads to a disagreement of the small area results with regard to the information already published for larger regions, making necessary the introduction of some constraints to benchmark the small area estimates to those provided by official agencies when aggregating the small areas. In this paper we develop an small area estimator by introducing some restrictions in a linear mixed model, in order to obtain fully concordant estimates with those provided using a synthetic BLUP at some aggregated level.

2 The Synthetic BLUP and the EBLUP

The synthetic BLUP is used to provide estimates at provincial level and it is derived from the superpopulation linear model

\[ y_{ij} = x_{ij}\beta + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2_{ij}c^{-1}_{ij}), \quad i = 1, \ldots, t, \quad j = 1, \ldots, N_i, \quad (1) \]

where in the \( i \)th area, \( y_{ij} \) is the \( j \)th value of the target variable, \( x_{ij} \) is the corresponding auxiliary information for the \( j \)th unit, \( \beta \) is the common slope and \( N_i \) is the number of elements with \( N = \sum_{i=1}^{t} N_i \). Here, \( t \) is the number of small areas, \( \epsilon_{ij} \sim N(0, \sigma^2_{ij}c^{-1}_{ij}) \) is the error term and \( c_{ij} = 1/x^2_{ij} \) are model weights to account for heteroscedasticity. Then the BLUP of the population mean in province \( h \), according to the prediction theory, is given by

\[ \hat{Y}^{syn}_h = f_h\bar{y}_{hs} + (1 - f_h)\hat{y}^{syn}_{hr}, \quad h = 1, \ldots, H, \quad (2) \]

where in province \( h \), \( f_h = n_h/N_h \) is the sampling fraction, \( n_h \) and \( N_h \) are the number of sampled and population units and \( \hat{y}^{syn}_{hr} \) is the BLUP corresponding to the non-sampled mean. The mean squared error (MSE) of predictor (2) is easy to calculate using the MSE definition. To obtain information at small area or county level, the following linear mixed model is used

\[ y_{ij} = x_{ij}\beta + v_i + \epsilon_{ij}, \quad i = 1, \ldots, t, \quad j = 1, \ldots, N_i, \quad (3) \]

where \( v_i \sim N(0, \sigma^2_v) \) is the county random effect, \( \epsilon_{ij} \sim N(0, \sigma^2_{ij}c^{-1}_{ij}) \) is the error term and \( c_{ij} = 1/x^2_{ij} \). The random effects and the random errors
are assumed to be independent. Then, the EBLUP of the county mean is expressed as

\[ \hat{\bar{Y}}_{i}^{Eblup} = f_i \hat{y}_{is} + (1 - f_i)\hat{Y}_{ir}^{Eblup}, \]  

(4)

where \( f_i = n_i/N_i \), \( n_i \) and \( N_i \) are the number of sampled and population units in the \( i \)th county and \( \hat{Y}_{ir}^{Eblup} \) is the EBLUP corresponding to the non-sampled mean. The MSE of the EBLUP (4) is a difficult task in small area estimation. Here, the results of Prasad and Rao (1990) are used.

### 3 The Restricted Predictor

The statistical offices are still reluctant to implement the EBLUP because there is a disagreement between the estimates already published for large regions and the results obtained when aggregating the small areas within those domains. To achieve the benchmarking property, we impose the following restrictions when estimating the parameters in Model (3),

\[ \sum_{i \in h} \hat{Y}_i^R = \hat{Y}_h^{Syn}, \quad h = 1, \ldots, H, \]  

(5)

where \( \hat{Y}_i^R \) and \( \hat{Y}_h^{Syn} \) are the modified and the synthetic BLUP predictors of the total in county \( i \) and province \( h \) respectively. Then, the restricted predictor is given by

\[ \hat{\bar{Y}}_{i}^{R} = f_i \hat{y}_{is} + (1 - f_i)\hat{Y}_{ir}^{R}, \]  

(6)

where \( \hat{Y}_{ir}^{R} \) is the modified predictor of the non-sampled mean in the \( i \)th area. Note that when adding the small area estimates within the same province, the value provided by the statistical office is attained. Moreover, this method can be seen as a tool to “borrow information” and to improve the estimates. The MSE of the restricted predictor (6) is provided combining closed expressions, similar to that obtained for the EBLUP, and different bootstrap terms (Pfffermann and Tiller, 2005).

### 4 Illustration and Results

The techniques are illustrated with real data from the 2002 Annual Business Survey of the Basque Country. Here, the goal is to provide total estimates of the gross sales by county and province for one specific industrial activity (not detailed for confidentiality) using as auxiliary information the number of employees in every establishment. The restricted predictor (6) is considered to obtain small area estimates fully concordant with those already published by the statistical office at provincial level using the synthetic BLUP (2). In addition, estimates derived from the EBLUP (4) are
TABLE 1. Provinces, sample and population sizes, total estimates, standard errors and coefficients of variation for the synthetic estimator

<table>
<thead>
<tr>
<th>Provinces</th>
<th>n_h</th>
<th>N_h</th>
<th>(\hat{Y}_{\text{Syn}})</th>
<th>s.e.((\hat{Y}_{\text{Syn}}))</th>
<th>c.v.((\hat{Y}_{\text{Syn}}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>P_1</td>
<td>7</td>
<td>103</td>
<td>35589.66</td>
<td>2820.40</td>
<td>0.08</td>
</tr>
<tr>
<td>P_2</td>
<td>17</td>
<td>375</td>
<td>146190.52</td>
<td>9127.58</td>
<td>0.06</td>
</tr>
<tr>
<td>P_3</td>
<td>13</td>
<td>333</td>
<td>120626.16</td>
<td>7661.15</td>
<td>0.06</td>
</tr>
<tr>
<td>Total</td>
<td>37</td>
<td>811</td>
<td>302406.34</td>
<td>18254.70</td>
<td>0.06</td>
</tr>
</tbody>
</table>

provided for comparison purposes. Table 1 shows sample and population sizes, the estimates, the standard errors and the coefficient of variation at provincial level and for the whole region (the total in the table) derived from the synthetic BLUP (2). Table 2 displays the total estimates obtained with the EBLUP and the restricted predictor. It also displays the bootstrap standard errors for both estimators obtained from 1000 bootstrap replications. Note that there is an increase in the standard errors of the restricted predictor with regard to the standard errors of the EBLUP. This is expected because the restricted predictor is not BLUP. However, the advantage of the proposed restricted predictor is that when adding the small area estimates within a province, the synthetic provincial estimate is obtained.

5 Discussion

Model-based methods, such as those based on linear mixed models, are used to derive estimators at county or municipality level. However, the estimates obtained with these techniques do not add up to the estimates obtained by statistical offices for a larger region composed of a set of small areas. In the example considered here, a synthetic estimator is used to provide estimates in each province by the statistical office and an EBLUP based on a linear mixed effects model is used to estimate by county. However, the aggregation of the county estimates within a province does not match up with the synthetic estimate. In this paper, we derive a predictor from a linear mixed model with restrictions to obtain the desirable concordance among the estimates. Its mean squared error is assessed by a bootstrap approximation because its direct calculation seems to be untractable. The method provides unique estimates for provinces avoiding a posterior adjustment or calibration.
restricted predictors

TABLE 2. Small areas or counties, sampled and population sizes, total EBLUP estimates, total restricted estimates, and bootstrap standard errors for the EBLUP and the restricted predictor

<table>
<thead>
<tr>
<th>Domain</th>
<th>(n_i)</th>
<th>(N_i)</th>
<th>(\hat{Y}_{EBLUP}^{Y})</th>
<th>(\hat{Y}^R)</th>
<th>s.e.((\hat{Y}_{EBLUP}^{Y}))</th>
<th>s.e.((\hat{Y}^R))</th>
</tr>
</thead>
<tbody>
<tr>
<td>A13</td>
<td>0</td>
<td>1</td>
<td>71.13</td>
<td>67.52</td>
<td>32.34</td>
<td>32.44</td>
</tr>
<tr>
<td>A15</td>
<td>0</td>
<td>5</td>
<td>2347.27</td>
<td>2230.42</td>
<td>454.97</td>
<td>464.17</td>
</tr>
<tr>
<td>A16</td>
<td>0</td>
<td>7</td>
<td>2845.18</td>
<td>2698.78</td>
<td>453.33</td>
<td>473.20</td>
</tr>
<tr>
<td>A26</td>
<td>0</td>
<td>45</td>
<td>18422.51</td>
<td>17604.61</td>
<td>1785.15</td>
<td>1973.42</td>
</tr>
<tr>
<td>A31</td>
<td>0</td>
<td>14</td>
<td>5690.35</td>
<td>5414.03</td>
<td>758.01</td>
<td>793.53</td>
</tr>
<tr>
<td>A33</td>
<td>0</td>
<td>47</td>
<td>18351.38</td>
<td>17392.11</td>
<td>1727.17</td>
<td>1860.73</td>
</tr>
<tr>
<td>A34</td>
<td>0</td>
<td>10</td>
<td>3983.25</td>
<td>3791.43</td>
<td>590.12</td>
<td>602.51</td>
</tr>
<tr>
<td>A35</td>
<td>0</td>
<td>4</td>
<td>640.16</td>
<td>609.34</td>
<td>162.59</td>
<td>164.26</td>
</tr>
<tr>
<td>A37</td>
<td>0</td>
<td>12</td>
<td>3912.12</td>
<td>3721.75</td>
<td>586.39</td>
<td>599.66</td>
</tr>
<tr>
<td>A11</td>
<td>1</td>
<td>5</td>
<td>1061.30</td>
<td>1017.74</td>
<td>214.76</td>
<td>229.25</td>
</tr>
<tr>
<td>A14</td>
<td>1</td>
<td>8</td>
<td>3609.69</td>
<td>3455.25</td>
<td>585.46</td>
<td>609.38</td>
</tr>
<tr>
<td>A21</td>
<td>1</td>
<td>21</td>
<td>6934.77</td>
<td>6683.24</td>
<td>865.55</td>
<td>920.42</td>
</tr>
<tr>
<td>A27</td>
<td>1</td>
<td>54</td>
<td>26236.38</td>
<td>25133.98</td>
<td>2250.28</td>
<td>2536.66</td>
</tr>
<tr>
<td>A23</td>
<td>2</td>
<td>21</td>
<td>9275.70</td>
<td>9075.59</td>
<td>881.21</td>
<td>929.52</td>
</tr>
<tr>
<td>A25</td>
<td>3</td>
<td>44</td>
<td>20706.12</td>
<td>19932.71</td>
<td>1909.16</td>
<td>2158.60</td>
</tr>
<tr>
<td>A22</td>
<td>4</td>
<td>67</td>
<td>22429.87</td>
<td>21500.37</td>
<td>2143.89</td>
<td>2432.74</td>
</tr>
<tr>
<td>A12</td>
<td>5</td>
<td>77</td>
<td>27503.71</td>
<td>26119.95</td>
<td>2120.30</td>
<td>2686.57</td>
</tr>
<tr>
<td>A24</td>
<td>6</td>
<td>123</td>
<td>47698.68</td>
<td>46260.02</td>
<td>3589.57</td>
<td>4268.69</td>
</tr>
<tr>
<td>A32</td>
<td>13</td>
<td>246</td>
<td>93257.80</td>
<td>89697.50</td>
<td>5779.57</td>
<td>6970.37</td>
</tr>
</tbody>
</table>

Acknowledgments: Special thanks to the Basque Country Statistical Office, EUSTAT, for providing the data. We gratefully acknowledge support from the Spanish Ministry of Science and Education (project MTM2005-00511)

References


Conditional AIC for nonlinear mixed effects models

Florin Vaida\textsuperscript{1} and Suzette Blanchard \textsuperscript{2}

\textsuperscript{1} Department of Family and Preventive Medicine, University of California at San Diego, La Jolla, CA 92093, USA; vaida@ucsd.edu
\textsuperscript{2} Department of Biostatistics, City of Hope, Duarte, CA 91010, USA; sblanchard@coh.org

\textbf{Abstract:} In this paper we propose a new model selection criterion for nonlinear mixed-effects model (NLME). The conditional AIC of Vaida and Blanchard (2005) is extended to NLME, using an appropriate definition of the effective degrees of freedom of the model, $\rho$. We show that $\rho$ is related to the effective degrees of freedom proposed for GLME by Lu, Hodges and Carlin (2006). The criterion approximates the conditional Akaike information and the goodness of the approximation depends on the degree of non-linearity of the model. The conditional AIC is useful when the purpose of the model is subject-specific rather than population-level prediction.

We use the criterion for model selection in the analysis of data from an ongoing international study of acute HIV infection.

\textbf{Keywords:} AIC; effective degrees of freedom; nonlinear mixed-effects models; primary HIV infection.

1 Introduction

Mixed effects models allow for population-level and subject-specific inference. As discussed by Spiegelhalter \textit{et al.} (2002) and Vaida and Blanchard (2005) the choice of model focus influences also model selection. Vaida and Blanchard (2005) show that the commonly used AIC is appropriate for population inference, but not for subject-specific inference. In the latter case they propose the conditional AIC (cAIC) and its theoretical counterpart, conditional Akaike Information (cAI). The cAIC was only computed for linear mixed effects models (LME). The purpose of the present paper is to extend cAIC to NLME and show its use in a situation of current importance in treatment of HIV infection.

The conditional AIC is based on the notion of effective degrees of freedom for a mixed effects model. This was put forth by Hodges and Sargent (2001) for LME and later extended to GLME by Lu, Hodges and Carlin (2006). They also showed that in practice $\rho$ gives comparable results to the Bayesian effective degrees of freedom based on the Deviance Information Criterion (DIC) of Spiegelhalter \textit{et al.}
2 Effective degrees of freedom for nonlinear mixed effects models

For each of \( m \) subjects, let \( y_i = (y_{ij})_{j=1 \ldots n_i} \) be the vector of responses for subject \( i \), e.g. HIV viral load, or drug concentrations, at time points \( t_{ij} \). In a general formulation the NLME is

\[
y_{ij} = f(\eta_{ij}) + \epsilon_{ij}
\]

where the conditional mean of \( y_{ij} \), \( f(\eta_{ij}) \) is non-linear in the linear predictor \( \eta_{ij} = x_{ij}\beta + z_{ij}b_i \); \( x_{ij} \) and \( z_{ij} \) are vectors (or more generally matrices) of covariates including draw times and dosing history, and \( \beta \) is a vector of population-level parameters. The subject-specific parameters \( b_i \) and error \( \epsilon_{ij} \) are modeled by

\[
b_i \overset{iid}{\sim} N(0, \sigma^2 D) \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)
\]

independent of each other. More general representations are possible and useful, including correlation structures for the error vector \( \epsilon_i = (\epsilon_{ij})_{j=1 \ldots n_i} \) and functional forms for \( \text{var}(\epsilon_i) \) including dependency of covariates (Pinheiro and Bates, 2000, chapter 4). The normal distribution assumption for the errors and/or random effects may be relaxed (Davidian and Giltinan, 1995).

A simpler case is the LME where the function \( f(\eta_{ij}) = \eta_{ij} \). A compact formulation of this is

\[
y = X\beta + Zb + \epsilon, \quad b \sim N(0, \sigma^2 D_0).
\]

In this LME the maximum likelihood estimator (MLE) \( (\hat{\beta}, \hat{b}) \) is given by the Henderson equations (Harville, 1977) which are equivalent formally to the weighted least squares equation from the linear model

\[
\begin{pmatrix}
y \\
0
\end{pmatrix} = M\begin{pmatrix}
\beta \\
b
\end{pmatrix} + \begin{pmatrix}
\epsilon \\
b
\end{pmatrix} = W^{-1}
\]

where

\[
M = \begin{pmatrix}
X & Z \\
0 & -I
\end{pmatrix} \quad \text{var}(\begin{pmatrix}
\epsilon \\
b
\end{pmatrix}) = \sigma^2 \begin{pmatrix}
I \\
0
\end{pmatrix} D_0 = W^{-1}
\]

with parameter \( \delta = (\beta^\top, b^\top)^\top \) and error \( \omega = (\epsilon^\top, b^\top)^\top \). From here we can write

\[
\hat{y} = \sigma^{-2}(X Z)(M^\top W M)^{-1}(X Z)^\top y = H_1 y,
\]

and define the effective degrees of freedom for the LME as \( \rho = \text{tr}(H_1) \) (Hodges and Sargent, 2001).

We extend \( \rho \) for NLME as follows. Using a linear expansion of the mean function, as in Linstrom and Bates (1990) we write \( f(\eta_{ij}) \approx f(\hat{\eta}_{ij}) \) +
\( f'(\hat{\eta}_{ij})(\eta_{ij} - \hat{\eta}_{ij}) \). Rearranging the terms in (1) we get \( w_{ij} = f'(\hat{\eta}_{ij})\eta_{ij} + \epsilon_{ij} \), or more explicitly

\[
\begin{align*}
    w_{ij} = u_{ij} \beta + v_{ij} b + \epsilon_{ij},
\end{align*}
\]

where \( u_{ij} = y_{ij} - f(\hat{\eta}_{ij}) + f'(\hat{\eta}_{ij})\tilde{\eta}_{ij} \), \( u_{ij} = f'(\hat{\eta}_{ij})x_{ij} \) and \( v_{ij} = f'(\hat{\eta}_{ij})z_{ij} \).

Treating formally (7) as a LME the effective degrees of freedom \( \rho \) are defined as in (6). This value depends on the set of \( f'(\hat{\eta}_{ij})'s \). To mark this dependency on the data we denote this by \( \hat{\rho} \). When \( \hat{\eta}_{ij} \) is replaced by \( \eta_{ij}^* = x_{ij}\beta^* + z_{ij}b^* \) in the definition of \( \rho \), we write this as \( \rho^* \), where \( \beta^* \) and \( b^* \) are the “true” parameter value and realized random effect.

An alternative approach was proposed for generalized LME by Lu, Hodges, and Carlin (2006), and it can be easily extended to NLME. We approximate the conditional log-likelihood of datum \( y_{ij} \), \( l(\eta_{ij}) \), using a second-order Taylor expansion around the MLE \( \hat{\eta}_{ij} \):

\[
    l(\eta_{ij}) \approx -\frac{1}{2\sigma_{ij}^2}(\xi_{ij} - \eta_{ij})^2 + \text{constant},
\]

where \( \xi_{ij} = \hat{\eta}_{ij} - l'(\hat{\eta}_{ij})/E\{l''(\hat{\eta}_{ij})\} \), and \( \sigma_{ij}^2 = -1/E\{l''(\hat{\eta}_{ij})\} \). Using the normal approximation from (8), \( \xi_{ij} \approx \eta_{ij} + \epsilon_{ij} \) with \( \epsilon_{ij} \sim N(0, \sigma_{ij}^2) \), we write formally, in analogy with (4)

\[
    \begin{pmatrix} \xi \\ 0 \end{pmatrix} = M \begin{pmatrix} \beta \\ b \end{pmatrix} + \begin{pmatrix} \epsilon \\ 0 \end{pmatrix}, \quad \text{var} \begin{pmatrix} \epsilon \\ b \end{pmatrix} = \begin{pmatrix} \Sigma & 0 \\ 0 & \sigma^2 \end{pmatrix} = W^{-1}
\]

where \( \Sigma = \text{diag}(\sigma_{ij}^{-2}) \) and \( M \) is given by (5). This suggests defining

\[
    \hat{\rho}_1 = \text{tr}(H_1), \quad H_1 = (X' Z)(M' WM)^{-1}(X' Z)' \Sigma^{-1}.
\]

Lu, Hodges, and Carlin use \( l''(\hat{\eta}_{ij}) \) instead of \( E\{l''(\hat{\eta}_{ij})\} \) in the definition of \( \xi_{ij} \) and \( \sigma_{ij}^2 \). We prefer taking expectations, since it eliminates a second order term and is equivalent to using Fisher scoring instead of Newton-Raphson in the maximization procedure. A similar approach was taken by Breslow and Clayton (1993) in computing the approximate MLE for GLME. Also, when \( \eta_{ij}^* \) is used instead of \( \hat{\eta}_{ij} \) in the definition of \( \sigma_{ij}^2 \) and \( H_1 \), we write \( \rho_1 = \text{tr}(H_1) \).

While in the latter approach the response is rescaled so that it is brought on the scale of \( \eta \) in the former the mean response \( f(\eta) \) is linearized in order to generate an approximately linear predictor \( f'(\hat{\eta})\eta \) on the scale of the response \( y \). The following result shows that the two approaches are equivalent.

**Proposition:** The two definitions of the effective degrees of freedom for the NLME are equivalent: \( \hat{\rho}_1 = \hat{\rho} \).

Moreover, in an asymptotic setting where the information about the random effects is accruing, as in Jiang, Jia and Chen (2001), \( \hat{\eta}_{ij} \) is consistent for \( \eta_{ij}^* \), and \( \hat{\rho} \) is a consistent estimator of \( \rho^* \).
3 Conditional AIC for nonlinear mixed-effects models

The conditional Akaike information (Vaida and Blanchard, 2005) is defined for a mixed-effects model as

\[
c_{\text{AI}} = -2E_{h(y,b^*)} E_{h(y_{\text{new}}|b^*)} \log g(y_{\text{new}}|\hat{\theta}(y), \hat{b}(y))
\]

(11)

where \( h \) is the operating model, i.e. the “true” density function generating each of the independent datasets \( y, y_{\text{new}} \), with the “true” random effect \( b^* \), and \( g \) is the candidate model, with \( \hat{\theta}, \hat{b} \) the MLE and empirical best linear unbiased predictor (EBLUP) for \( \theta = (\beta, \sigma^2) \) and random effects \( b \). The main result follows:

**Theorem 1:** If the operating and candidate models are from the NLME family (1), (2), and \( \sigma^2 \) and \( D \) is known, the conditional AIC is given by

\[
c_{\text{AIC}} = -2 \log g(y|\hat{\beta}(y), \hat{b}(y)) + 2K
\]

(12)

where \( K = \rho \), i.e. \( \rho \) counts the effective number of parameters in the model \( g \). Under regularity conditions \( c_{\text{AIC}} \) is asymptotically unbiased as an estimator of \( c_{\text{AI}} \).

The following result covers the case for \( \sigma^2 \) unknown:

**Theorem 2:** Under same setting as in Theorem 1, but with \( \sigma^2 \) unknown and \( D \) known, the conditional AIC given by (12) with \( K = \hat{\rho} + 1 \) is asymptotically unbiased estimator of \( c_{\text{AIC}} \). A small-sample adjustment can be included as \( K = N/(N - p - 2)(\hat{\rho} + 1) \).

4 A simulation study

The results of Vaida and Blanchard (2005) suggest that when \( D \) is unknown its influence is asymptotically negligible. However, an adjustment needs to be made in practice, although it is not as large as the number of parameters in \( D \). We performed a simulation study in which \( c_{\text{AI}} \) was compared to \( E(c_{\text{AIC}}) \). We considered a pharmacokinetic one-compartment model with 10 clusters of \( m \) observations each, of values in the interval \((0,24)\). We took combinations of 3 values for \( m \) and 3 values for the residual standard deviation \( \sigma \), while keeping \( D \) constant. The \( c_{\text{AIC}} \) bias is reported in Table 1. As expected, we found that the bias reduces with increasing sample size and reduced residual standard deviation \( \sigma \). The bias includes both effects of unknown \( D \) and model non-linearity.
5 An HIV study

The AIEDRP CORE study is an international observational study of primary HIV infection. For this analysis we are focusing on 320 subjects with acute HIV infection (presenting for enrollment within approximately 24 days of HIV infection), and we are analyzing their HIV RNA viral load prior to starting anti-retroviral treatment.

In acute HIV infection the HIV RNA viral load shows an early spike, corresponding to the acute stage, followed by a gradual decrease to a stochastic steady state. This setpoint value may differ between individuals, and is of central interest here. The viral setpoint characterizes the severity of infection, it may relate to the strength of the subject’s immune system, it may predict clinical progression of the disease, and it determines the timing of starting antiretroviral treatment. We fitted several models to these data, including a four-parameter logistic model

\[ y_{ij} = \alpha_{1i} + \alpha_2 [1 + \exp \left( (t_{ij} - \alpha_3) / \alpha_4 \right)]^{-1} + \epsilon_{ij} \]

where \( y_{ij} \) is the log_{10} HIV RNA for subject \( i \) at time \( t_{ij} \). The parameters \( \alpha_{1i} \) and \( \alpha_2 \) are the setpoint value and the decrease from the maximum HIV RNA; \( \alpha_4 \) is a scale parameter modelling the rate of decline, and \( \alpha_3 \) is a location parameter indicating the time of achieving the HIV RNA mid-point value. In order to force the parameters to be positive we reparametrized the model to \( \beta_{1i} = \log(\alpha_{1i}) \), \( \beta_k = \log(\alpha_k) \), \( k = 2, 3, 4 \). The setpoint \( \alpha_{1i} \) was taken to be random: \( \beta_{1i} = \beta_1 + b_i, b_i \sim N(0, \sigma_{b1}^2) \). The choice of final model is based on the conditional AIC.

References


Using profile likelihood for model selection with application to proportional hazards mixed models

Ronghui Xu¹, Florin Vaida¹ and David P. Harrington²

¹ Division of Biostatistics and Bioinformatics, Department of Family and Preventive Medicine, University of California, San Diego
² Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health

Abstract: Proportional hazards mixed effects model (PHMM) was recently proposed, which incorporates general random effects of arbitrary covariates and includes the frailty model as a special case. In this paper we consider selection of both nested and non-nested PHMMs. Using the profile likelihood we can define both a likelihood ratio statistic and an Akaike information for general models with nuisance parameters. Asymptotic quadratic expansion of the log profile likelihood allows derivation of the asymptotic null distribution of the likelihood ratio statistic including the boundary cases, as well as unbiased estimation of the Akaike information by an Akaike information criterion. For computation of the likelihood under PHMM we apply three algorithms: Laplace approximation, reciprocal importance sampling and bridge sampling. We compare the three algorithms under different data structures, and apply the methods to a multi-center lung cancer clinical trial.

Keywords: Akaike information; likelihood ratio test; testing on the boundary; reciprocal importance sampling; bridge sampling.

Vaida and Xu (2000) studied the proportional hazards model with mixed effects (PHMM). It includes the more classical ‘frailty’ models with random effects on the baseline hazard, but also allows random covariate effects. In this way it is able to model covariate by cluster interactions, such as varying treatment effects in a multi-center clinical trial. The model is of the form

\[ \lambda_{ij}(t) = \lambda_0(t) \exp(\beta' Z_{ij} + b_i' W_{ij}), \]

where \( \lambda_{ij}(t) \) is the hazard function of the \( j \)-th observation from the \( i \)-th cluster, \( b_i \) is a vector of random effects for the \( i \)-th cluster, and \( Z_{ij}, W_{ij} \) are the covariate vectors for the fixed and random effects. This model contains a multivariate random effect with arbitrary design matrix in the log relative risk, in a way similar to the linear, generalized linear and nonlin-
ear mixed models. Vaida and Xu developed the nonparametric maximum likelihood estimator of the parameters in this model, computed using the EM algorithm and Markov Chain Monte Carlo (MCMC) methods. Assume that the data consist of possibly censored event time observations from \( n \) clusters, with \( n_i \) observations in each cluster, \( i = 1 \ldots n \). Within a cluster the observations are dependent, but conditional on the cluster-specific \( d \times 1 \) vector of random effects \( b_i \), the survival times \( T_{ij} \) are independent and follow the proportional hazards model (1). In (1) \( W_{ij} \) is often a subset of \( Z_{ij} \), apart from possibly a ‘1’ which represents the cluster effect on the baseline hazard. To insure identifiability, we assume that \( E(b_i) = 0 \). For distribution of the random effects we also assume that

\[
b_i \overset{iid}{\sim} N(0, \Sigma)
\]

(2)
as in Vaida and Xu (2000). Note that the other commonly used frailty distribution, the gamma distribution, is not suitable here. This is because it is not scale-invariant so that the inference is not invariant under a change of measuring unit for the random effects.

The data from subject \( j \) in cluster \( i \) can be written \( y_{ij} = (X_{ij}, \delta_{ij}, Z_{ij}, W_{ij}) \), where \( X_{ij} \) is the possibly right-censored failure time and \( \delta_{ij} \) is the failure-event indicator. Let \( y_i = (y_{i1}, \ldots, y_{in_i}) \) be the data for cluster \( i \). For cluster \( i \), conditional on the random effect \( b_i \), the log-likelihood is

\[
l_i = l_i(y_{ij}; \beta, \lambda_0 | b_i) = \sum_{j=1}^{n_i} \{ \delta_{ij} \log \lambda_0(\beta'Z_{ij} + b_i'W_{ij}) - \lambda_0(X_{ij})e^{\beta'Z_{ij} + b_i'W_{ij}} \},
\]

(3)

where \( \Lambda_0(t) = \int_0^t \lambda_0(s) \, ds \). We rewrite the parameter for the baseline hazard in the following as \( \lambda \), to be consistent with the general semiparametric model framework that we will use. The likelihood of the observed data is then

\[
L(\theta) = \prod_{i=1}^{n} \int \exp(l_i(b_i; \Sigma)) \, db_i,
\]

(4)

where \( \theta = (\beta, \Sigma, \lambda) \), and \( p(\cdot; \Sigma) \) is the multivariate normal density. Usually no closed-form expression is available for \( L(\theta) \) and its calculation involves \( d \)-dimensional integration.

In this paper we consider both the likelihood ratio test and an Akaike information criterion (AIC) for model selection in the presence of nuisance parameters. They turn out to be derived from the same asymptotic expansion of a log profile likelihood. They also share the same computational algorithm, which we develop under PHMM. Under PHMM our parameter of interest is \( \phi = (\beta, \Sigma) \), whereas \( \lambda \) is seen as a nuisance parameter. The log-profile likelihood for \( \phi \) is

\[
pl(\phi) = \sup_{\lambda} l(\phi, \lambda),
\]

(5)
where \( l(\phi, \lambda) = \log L(\theta) = \sum l_i \) is the log-likelihood. Following Murphy and van der Vaart (2000), under the conditions stated in their Theorem 1, the log profile likelihood behaves as a quadratic function asymptotically; i.e. for any random sequence \( \phi_n \) such that \( \| \phi_n - \phi_0 \| = O_p(1/\sqrt{n}) \) where \( \phi_0 \) is the true parameter value,

\[
\frac{1}{n} \{ l(\phi_n) - l(\phi_0) \} = (\phi_n - \phi_0)' A - \frac{1}{2} (\phi_n - \phi_0)' I (\phi_n - \phi_0) + o_p \left( \frac{1}{n} \right),
\]

where \( A = \sum_1^n s(y_i)/n \), \( s \) is the efficient score for \( \phi \), i.e. the ordinary observed score function minus its orthogonal projection onto the closed linear span of the score functions for the nuisance parameter \( \lambda \), and \( I \), its covariance matrix, is the efficient Fisher information matrix (Murphy and van der Vaart, 2000; Severini and Wong, 92).

Using the profile likelihood we can define both a likelihood ratio statistic and an Akaike information for general models with nuisance parameters. Asymptotic quadratic expansion (6) of the log profile likelihood allows derivation of the asymptotic null distribution of the likelihood ratio statistic including the boundary cases, as well as unbiased estimation of the Akaike information by an Akaike information criterion. For computation of the likelihood under PHMM we apply three algorithms which are increasingly intensive computationally: Laplace approximation, reciprocal importance sampling and bridge sampling. We compare the three algorithms under different data structures. In addition we show that in finite samples the asymptotic null distribution of the likelihood ratio statistic is inaccurate for testing zero variance components, while the recent results of Crainiceanu and Ruppert (2004) appear to hold empirically. In the following we apply the methods to a multi-center lung cancer clinical trial which has been previously noted to have heterogeneous effects of treatment and other covariates among the centers.

1 An example

In this section we consider the multi-center non-small cell lung cancer trial that was used as an example in Vaida and Xu (2000). The trial enrolled 579 patients from 31 institutions. The primary endpoint was patient death. There were two randomized treatment arms in the trial, a standard chemotherapy (CAV) arm and an alternating regimen (CAV-HEM) arm. Other important covariates that affected patient survival were: presence or absence of bone metastases, presence or absence of liver metastases, performance status at study entry and whether there was weight loss prior to entry. Gray (1995) used a score test for the existence of random treatment effect, and found it to be significant.

In the following we consider the three nested models of Vaida and Xu (2000). They all include the fixed effects of the five covariates. Model 1 includes no random effect; Model 2 includes a random treatment effect; and
Model 3 includes random treatment and random bone metastases effects. The estimate of the other variance components corresponding to potential random effects for the rest three of the covariates, as well as random center effect on the baseline hazard function (see also Gray, 1995), converged to zero during the EM algorithm (Vaida and Xu, 2000). The parameter estimates under the three models were given in Table 1 of Vaida and Xu (2000). Table 1 here gives minus twice the log likelihood values for the three models, computed using Laplace approximation, reciprocal importance sampling and bridge sampling for models 2 and 3. Note that the likelihood for Model 1 can be computed directly since it is the usual Cox model. The likelihood values for all three models are computed after 50 EM steps where the maximum likelihood estimate has converged; the sample sizes for Gibbs sampler during MCEM are 100 initially and increased to 1000 for the last 10 EM steps. The Monte Carlo sample sizes for RIS and BS are 1000, respectively. From the table we see that the values of the log likelihoods agree well among the three computational methods.

As seen in the table, if we are to test Model 2 versus Model 1 using the likelihood ratio statistic, its sampling distribution under Model 1 is asymptotically \((\chi^2_0 + \chi^2_1)/2\) with critical value of 2.71 at .05 significance level. Model 1 is then rejected in favor of Model 2. Similarly, to test Model 3 versus Model 2, the likelihood ratio statistic is again asymptotically \((\chi^2_0 + \chi^2_1)/2\) under Model 2, and Model 2 is rejected in favor of Model 3. Note that the finite sample distribution we considered in Section 5 puts more point mass at zero, leading to even smaller critical values for the likelihood ratio statistic.

Alternatively, we can use the profile AIC to compare the nested models. From the table it also clear that the larger models are chosen by the criterion.

<table>
<thead>
<tr>
<th>Model</th>
<th>Laplace</th>
<th>RIS</th>
<th>BS</th>
<th>pAIC*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1*</td>
<td>7232.80</td>
<td>7232.80</td>
<td>7232.80</td>
<td>7242.80</td>
</tr>
<tr>
<td>2</td>
<td>7228.98 (3.82)</td>
<td>7228.80 (4.00)</td>
<td>7228.78 (4.02)</td>
<td>7240.80</td>
</tr>
<tr>
<td>3</td>
<td>7222.72 (6.26)</td>
<td>7222.55 (6.25)</td>
<td>7222.60 (6.18)</td>
<td>7236.55</td>
</tr>
</tbody>
</table>

RIS - reciprocal importance sampling, BS - bridge sampling.
* likelihood computed directly for Model 1.
In (·) are the likelihood ratio statistics between the model and its immediate submodel (3 vs. 2; 2 vs. 1).
2 Discussion

In this paper, motivated by model selection problems under PHMM, we developed the profile likelihood ratio test and a profile Akaike information criterion that are generally applicable to models with nuisance parameters. The development was based on the asymptotic quadratic expansion of the profile likelihood function. The profile likelihood ratio test for the null hypothesis that lies in the interior of the parameter space was given in Murphy and van der Vaart (2000); here we further developed it for testing on the boundary. The profile \( \text{AIC} \) has not been previously proposed in the literature, to our best knowledge. It applies to both parametric and semiparametric models, where for the latter models the focus is on the finite dimensional parameter. The profile \( \text{AIC} \) provides a theoretical justification for the use of the partial likelihood in the \( \text{AIC} \) under the classic Cox model.

Model selection has been an area of growing interest in the recent years. In this paper we restricted our attention to the classic derivation of the Akaike information criterion. However we acknowledge, as Longford (2005) pointed out, that, whatever the selection criterion, single-model based inference can be inherently biased. Alternatives may include the use of a mixture of plausible models, and the focused information criteria of Claeskens and Hjort (2003). The associated new challenges of such improvements in practice are model interpretability and variability of inferences following the model averaging or selection.

The asymptotic quadratic expansion of the profile likelihood requires certain conditions to be met, such as those given in Murphy and van der Vaart (2000). Alternatively, asymptotic normality of the \( \text{MLE} \) also implies that the likelihood surface is asymptotically quadratic near the true parameter values, which in turn implies that the same holds for the profile likelihood. The asymptotic properties of the \( \text{MLE} \) have been established for the gamma frailty models (Murphy, 1994, 1995; Parner, 1998), and empirical evidence supports the same for the PHMM (Maple et al., 2002). Rigorous proof of the asymptotic normality of the \( \text{MLE} \) under PHMM is done in a separate project by the first author and colleagues.

For computation of the maximized likelihood, the Laplace approximation is the most straightforward but is only accurate when the cluster sizes are reasonably large. In view of the \( \text{MCEM} \) algorithm that is used to fit the PHMM, the additional computation of \( \text{RIS} \) or \( \text{BS} \) is often comparable to one step of the \( \text{MCEM} \). Therefore we include \( \text{RIS} \) and \( \text{BS} \) as default in our computational program.

Finally, under linear mixed models when the interest lies in the inference of the random effects themselves, Vaida and Blanchard (2005) propose a conditional \( \text{AIC} \) using the notion of effective degrees of freedom. The usefulness of conditional inference carries over to PHMM, and it is our future work to develop a conditional \( \text{AIC} \) under the PHMM. Additionally,
the finite sample distribution of the likelihood ratio statistic for testing zero variance components is another area that requires further work.

References


Author Index

Adamska, E, 252
Aerts, M, 346
Aguilar-Saborit, J, 393
Amaral Turkman, M, 65
Armero, C, 70, 369

Böckenholt, U, 119
Bailey, TC, 74
Barry, D, 336
Barry, S, 82
Berridge, D, 193
Bijnens, L, 346
Blanchard, S, 499
Blanco, C, 90
Blatná, D, 482
Bowman, A, 82
Brown, J, 438
Brown, PJ, 373
Burnett, R, 90

Cakmak, S, 90
Calle, ML, 94
Cappelli, C, 438
Carrasco, L, 159
Castellanos, ME, 369
Castillo, J, 102
Castillo-Morales, A, 164
Cazzaro, M, 107
Cecere, S, 111
Claeskens, G, 3
Colombi, R, 107
Condon, J, 271
Connolly, J, 216
Crainiceanu, M, 303
Cruyff, MJLF, 119
Currie, I, 143, 279
Dahyot, R, 127
Daoudi, J, 102

Dasu, T, 131
Debusho, LK, 138
Durban, M, 143

Eilers, PHC, 143, 151, 313
Einbeck, J, 474
Escaramis, G, 159
Escarela, G, 164
Espinal, A, 94

Faddy, MJ, 172
Faes, C, 346
Farnandez, M, 90
Fenlon, JS, 172
Finselbach, HK, 182
Firth, D, 190
Fitzmaurice, GM, 14
Fried, R, 369
Friedl, H, 401

Ganjali, M, 193
Georganta, Z, 257
Geys, H, 346
Gilchrist, R, 234
Goicoa, T, 363, 494
Gonçalves, L, 65
González-Ramajo, B, 363
Greven, S, 201
Griffin, JE, 373
Gustafson, P, 340
Gómez, G, 94

Ha, ID, 208
Haines, LM, 138
Harrington, DP, 505
Haslett, J, 26, 448
Hawkins, MJ, 216
Heiner, K, 222
Held, L, 38
Heller, G, 226
Hennerfeind, A, 286
Hinde, J, 222, 474
Hrdličková, Z, 242
Hsu, C-H, 246
Hyde, BP, 216
Jacqmin-Gadda, H, 422
Jara, A, 111
Jordan, C, 330
Küchenhoff, H, 201
Kaczmarek, Z, 252
Kandiloro, H, 257
Kauermann, G, 303
Kazemi, I, 263
Kelly, GE, 271
Kirkby, J, 279
Kneib, T, 286
Kosmidis, I, 294
Krivobokova, T, 303
Kulinskaya, E, 309
López-Quílez, A, 70
López-Sánchez, R, 70
López-Sagaseta, M, 363
Lambert, J, 313
Lang, JB, 321
Larriba-Pey, J-L, 393
Lee, Y, 208
Lesaffre, E, 111
Li, W, 90
Livada, A, 257
Livingstone, V, 330
Lynch, C, 336
MacKenzie, G, 208
MacNab, YC, 340
Maringwa, JT, 346
Marx, BD, 151
Mayoral, AM, 369
McKinley, TJ, 74
Mena, H, 164
Mendes, JM, 354
Militino, AF, 363, 494
Molenberghs, G, 346
Morales, J, 369
Morgan, BJT, 50
Morris, D, 373, 380
Muggeo, VMR, 389
Muntés-Mulero, V, 393
Neubauer, G, 401
O’Kelly, M, 409
Ombredane, D, 380
Oskrochi, G, 414
Pérez-Cassany, M, 393
Parnell, A, 26
Pavlík, T, 466
Pelletier, J, 131
Penn, R, 193
Peters, A, 201
Proust, C, 422
Puig, P, 427
Purutçuoğlu, V, 431
Rea, W, 438
Reale, M, 438
Reed, WJ, 442
Rigby, B, 226
Ryan, M, 216
Salter-Townshend, M, 26, 448
Scarrott, C, 457
Schimek, MG, 466
Schulte, RPO, 216
Shamley, D, 414
Shkedy, Z, 346
Sofroniou, N, 474
Stasinopoulos, M, 226
Staudte, G, 309
Tawn, J, 457
Teisman, A, 346
Tresl, J, 482
Tunnicliffe-Wilson, G, 457
Turkan, KF, 354
Turner, HL, 190
Tutz, G, 486
Ugarte, MD, 363, 494
Ulbricht, U, 486
Vaida, F, 499, 505
van Ammel, K, 346
van den Hout, A, 119, 234
van der Heijden, PGM, 119, 234
Watkins, AJ, 182
Wit, E, 431
Wright, J, 131
Xu, R, 505
Zé-Zé, L, 65