Proceedings of the 25th International Workshop on Statistical Modelling

July 5-9, 2010

Glasgow

Adrian W. Bowman
(editor)

Proceedings of the 25th International Workshop on Statistical Modelling.
Glasgow, July 5-9, 2010
Adrian W Bowman, editor
Glasgow 2010.

Editor:
Adrian Bowman
Department of Statistics
The University of Glasgow
Glasgow G12 8QQ Scotland, UK
adrian@stats.gla.ac.uk

Printed by The University of Glasgow Print Unit

Scientific Programme Committee

- Carmen Armero
  University of Valencia, Spain
- Jim Booth
  Cornell University, Ithaca, USA
- Adrian Bowman (Chair)
  University of Glasgow, UK
- Mark Brewer
  Biomathematics and Statistics Scotland (BIOSS), UK
- Rosa Crujeiras
  University of Santiago de Compostela, Spain
- Jutta Gampe
  Max Planck Institute for Demographic Research, Germany
- Leo Held
  University of Zurich, Switzerland
- Robin Henderson
  University of Newcastle, UK
- John Newell
  National University of Ireland, Galway, Ireland
- Tony Pettitt
  Queensland University of Technology, Australia
- Geert Verbeke
  K.U. Leuven, Belgium

Preface

In 2010, the International Workshop on Statistical Modelling celebrates its 25th birthday. Twenty-five years old is a good age to be, as it combines maturity with vitality, and both of these attributes are evident in abundance in this year's workshop. The invited papers demonstrate both the current exciting developments in methodology and the key role which statistical modelling plays in a very wide variety of modern applications. This is amplified in both breadth and depth in the large number of contributed papers at this year's workshop. Some aspects of the programme honour the past and celebrate the origins and development of statistics, as befits a birthday event. However, very importantly, there is a major focus on the future, marked in particular by a substantial number of student contributions. It is good to see that statistical modelling is in very good health.

This year the IWSM comes to Glasgow, a city with a rich history and a dynamic culture. We hope that you will take full advantage of your stay here to enjoy the character of the city and a variety of social events on the workshop programme aim to help you in doing this. Glasgow is also located close to some spectacular Scottish scenery and we hope that you will have time to sample a little of this.

However, Glasgow also has a rich history in statistics and the Department of Statistics in the University of Glasgow is particularly pleased to be able to host the IWSM. The aims of the workshop coincide with the aims of the Department, to promote the subject of statistics and its application to important scientific problems, and so we look forward very much to meeting you and interacting with you.

The scientific success of the workshop depends on the participants, although this is focussed by the scientific committee whose contributions are very much appreciated. The organisational success of the workshop necessarily depends on a much smaller number of people, and the pivotal role is played by our local organiser, Claire Ferguson. With the very able assistance of Sarah Barry, plus a variety of administrative and IT staff, Claire has put very considerable time, energy and expertise into the preparations for the workshop. We hope that you enjoy the fruits of these labours. For my own part, I would also like to thank Ludger Evers for invaluable assistance in the technicalities of preparing the conference proceedings.

So welcome to Glasgow, and enjoy the workshop!

Adrian Bowman
Glasgow, July 2010

Part 1. Invited papers

Part 2. Contributed papers

Part 1. Invited papers

The past and future of statistical modelling

Murray Aitkin[1]

[1] Department of Mathematics and Statistics, University of Melbourne, Australia 3000

Abstract: This paper reviews advances in statistical modelling since 1985, as reflected in IWSM proceedings, and comments on corresponding changes in inferential approaches. The future development of modelling, and the inferential tools needed for analysis, are conjectured.

Keywords: Modelling, inference, Bayesian analysis, model comparisons

1 Introduction

The 25 years since 1985 have seen remarkable advances in the power of statistical modelling, and in the complexity of statistical inference.
By 1985 it was already clear that the GLIM IWLS approach to maximum likelihood in the exponential family could be greatly extended, by several methods. Continuous or finite mixtures of the exponential family could be handled through the EM algorithm, and non-exponential family distributions could be handled by extending IWLS by iteratively updating any needed function appearing in the adjusted dependent variable, score or information. In both these generalizations, it was possible to lose both the unimodality of the likelihood and the monotone and rapid convergence of the scoring algorithm.
It was also already clear that the Neyman-Pearson theory had difficulty with some properties of the exponential family and these extensions, particularly finite mixtures from latent class models. Models with two levels of latency required nested EM algorithms and complex calculations for the information matrix; the resulting standard errors were of unknown validity. Testing for the number of components in the mixture presented apparently insuperable problems for the distribution of the likelihood ratio test statistic. But even simpler problems involving (for example) ratios of parameters could not be handled by delta methods, and profile likelihoods were always of overstated precision.
Extensions of classical theory through saddlepoint methods and the p^* formula of Barndorff-Nielsen provided solutions to some of these problems, and quasi-likelihoods gave approximate solutions to others, though the properties of the latter solutions had to be investigated on a case-by-case basis.

2 Bayesian developments

Fully Bayesian extensions of the EM algorithm were not discussed in the Dempster, Laird and Rubin paper of 1977, beyond their application to empirical Bayes estimation through posterior modes, though the use of multiple imputation of missing data was discussed extensively in the books by Little and Rubin (1987) and Rubin (1987). The 1987 JASA paper by Tanner and Wong set out the stochastic version of the EM algorithm, called the Data Augmentation algorithm. It used the same "complete data" formulation as EM, in which the observed data are augmented by "latent data" to give the complete data, but alternated between *posterior draws* from the conditional posterior of the latent data given the current parameter draws – multiple imputations of the latent data – and from the conditional posterior of the parameters, given the current latent data draws. The priors for the model parameters could be flat or non-informative for both the model parameters and the latent data. Convergence was more difficult to assess than for EM, since it was in *distribution* rather than to a local point maximum.

This development solved the problem of standard errors for maximum likelihood parameter estimates, by generating the *full posterior distributions* through simulations, and provided full posteriors for *any parametric function* in addition, solving another problem in the frequentist framework. The subsequent book by Tanner (first edition 1991) gave a detailed exposition of this general approach, and in the 3rd edition 1996 extended the discussion to reflect developments in Markov chain Monte Carlo analysis.

3 Model comparisons

In repeated-sampling theory, the comparison of complex nested models had to rely on the likelihood ratio test, and/or on bootstrapping. Neither of these was adequate, and sometimes not even relevant, for the increasingly complex models which could be handled by MCMC, and even obtaining the maximized likelihood was increasingly difficult. For non-nested models, the optimality of the Cox approach (given in Pace and Salvan 1990) was limited, and required an arbitrary designation of null and alternative models for the sampling distribution of the test statistic.

Where maximized likelihoods could be obtained, AIC and BIC were increasingly used, despite substantial sampling studies which showed unsatisfactory performance of both methods with increasing sample size: AIC favoured over-complex models, BIC favoured under-complex models. Users generally computed both and hoped that they agreed. Since these are *decision* criteria, the issue of the *strength of evidence* provided by either, for one model over another, was unclear.

For Bayesians, model comparisons for either nested or non-nested models generally used the Bayes factor, following Jeffreys (1961). The major

difficulty was with its computation, and its dependence on the hyper-parameters of the proper prior needed. The hyper-parameters needed to be set for MCMC so that the initial draws from the prior were from a region of appreciable likelihood, so that the chain did not stay trapped in a region of flat likelihood. These hyper-parameter values however also determined the value of the integrated likelihood, which could lead to wildly diverse values of the integrated likelihoods for different sets of hyper-parameter values.
An outstanding example of this difficulty in finite mixture distributions was given by Aitkin (2001) in the first issue of Statistical Modelling, for the Postman et al (1987) data on recession velocities of 82 galaxies, analysed by Roeder (1990) and many other Bayesians, and shown below in a stem-and-leaf plot:

```
                              99
                              97
                              92
                              91
                              86
                              86 99
                              85 88
                              66 85    91
                              55 82    89
                              54 80    75 71
                              53 63    75 71
                              47 42    50 67
                              44 22 96 37 54 99
78                            34 22 92 31 54 72
56                         93 34 20 81 25 48 37
48                         60 33 18 70 24 26 29
35 41                17    55 07 18 49 21 24 29    99                79
17 23                08    42 05 17 14 19 21 13 63 96                07    28
-------------------------------------------------------------------------------
 9.10.11.12.13.14.15.16.17.18.19.20.21.22.23.24.25.26.27.28.29.30.31.32.33.34.

                                  velocity
```

The posterior distribution for the number K of components in a normal mixture distribution for velocity varied, across different Bayesian analyses, from a spike at 3, through a spike at 4, to very diffuse distributions over the range 3-13, with a mode at 6 or 7:

K	3	4	5	6	7	8	9	10	11	12	13
EW1	-	.03	.11	.22	.26	.20	.11	.05	.02	-	-
EW2	.02	.05	.14	.21	.21	.16	.11	.06	.03	.01	-
CC1	.64	.36	-	-	-	-	-	-	-	-	-
CC2	.004	.996	-	-	-	-	-	-	-	-	-
PS	-	-	-	.03	.39	.32	.22	.04	-	-	-
RW	.999	.001	-	-	-	-	-	-	-	-	-
RG	.06	.13	.18	.20	.16	.11	.07	.04	.02	.01	.01
N	.02	.13	.16	.25	.20	.13	.06	.03	.01	.01	-

When asked about this, many Bayesians have argued that *of course*, this results from the different priors used by the different analysts. This only accentuates the difficulty of the integrated likelihood approach, and raises major concerns about the validity of model choices, and model averaging, by Bayes factors in practical applications. This is not a theoretical issue: an example of the complete reversal of conclusions about the best supported model, from Bayes factors and *any other* model comparison approach, frequentist or Bayesian, is given in Liu and Aitkin (2008).

4 The current state of statistical modelling

The applications of statistical modelling are now so wide that it is difficult to give an overview. I focus on two major developments in modelling which have become generally popular, despite difficulties:

- the replacement of fully parametric regressions by partly parametric and partly nonparametric regressions using *smooth* functions determined through cubic splines or other semi-parametric methods;
- the use of spatio-temporal two-level models with parametrized covariance structure, commonly involving nearest neighbours.

The connection between penalized splines and variance component models has been elegantly exploited to provide a straightforward form of smoothing of departures from the parametric regression, using random effects attached to the positive part of the quadratic at each chosen knot. The degree of smoothing is based on the MLE (or posterior mode) of the variance component penalty; this leaves the nonparametric "smooth" function wiggly in places as all knots have random effects whose posterior means (BLUP estimates) contribute to the wiggles. It is difficult to determine which wiggles are meaningful and interpretable; it would be possible, though tedious, to remove non-important wiggles by sequentially deleting knots at which the random effects were small.
The introduction of general spatial covariance between areas in a two-level model greatly complicates the likelihood and requires pseudo-likelihoods

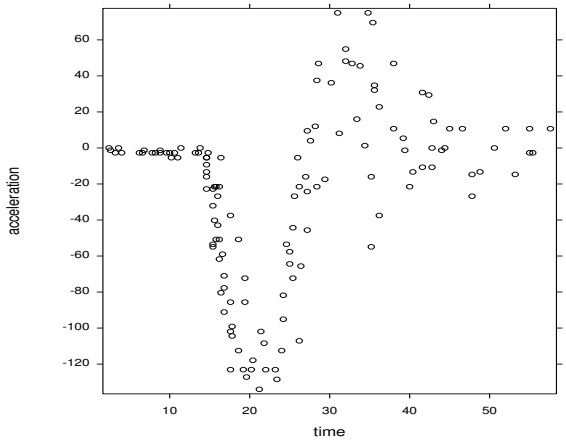


FIGURE 1. Motorcycle helmet acceleration

for practical computation; these allow relatively straightforward MCMC analysis, but imply conditional independence of area random effects given the neighbour random effects, which is generally invalid. The effect of using the pseudo-likelihood, in place of the true likelihood, on parameter estimates and standard errors is very difficult to determine. It also makes model comparisons difficult between the pseudo-likelihood for the spatial dependence model and the correct likelihood for the spatial independence model.

5 A framework for smooth modelling

The motorcycle helmet acceleration data set from Silverman (1985) in Figure 1 provides a challenging example for modelling. The fully nonparametric model used by Silverman, for observations y_i at design points t_i, is

$$y_i = g(t_i) + \epsilon_i,$$

where the ϵ_i are uncorrelated with zero mean, and g is a "smooth" function. Without some restriction on the ϵ_i, the model is not well-defined; we assume that $\epsilon_i \sim N(0, \sigma^2(t_i))$, the variance function being allowed to vary across time in some smooth way (observation-specific variances would not be identifiable). The combination of different phases after the simulated impact, together with very different variabilities in these phases, requires a quite complex model. We propose the *mixture of experts* model, popular in computer science applications (Jacobs et al 1991).

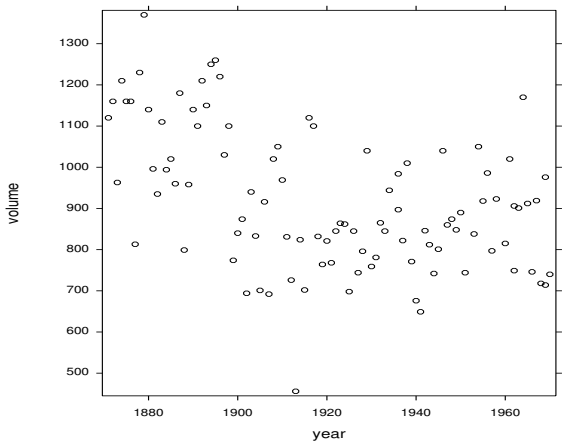


FIGURE 2. Nile flood volumes

5.1 Mixture of experts model

This is a finite mixture-of-regressions model, each component of the mixture being an "expert" (a confusing computer science term) with different regressions in each component on the explanatory variables for location (and for variability, if needed). The constant mixing proportion π_k for the k-th component in the simple finite mixture can be replaced if necessary by a different logistic regression for each k on the explanatory variables. This provides a smooth transition from one phase to the next; the necessary complexity of the various regressions can be assessed by the usual frequentist or Bayesian approaches.

For the motorcycle data, there appear to be five phases: an initial constant mean with very low variance before the impact, a rapid decline and rebound, with slowly increasing variance – this may be one or two phases – a slower decline with much larger variance and a nearly constant final phase with intermediate variance.

A simpler example is of the location of the changepoint in the volume of the annual Nile flood. The flood volume is shown for 100 years in Figure 2. A changepoint is to be estimated, using a normal model with different means but the same variance before and after the changepoint. This is far from clearly defined, but is estimated, by various methods, at 1898. A more natural model is a mixture of normals with different means; this allows the change point to be uncertain over a range. An unconstrained mixture allows for the possibility of several changes beween states.

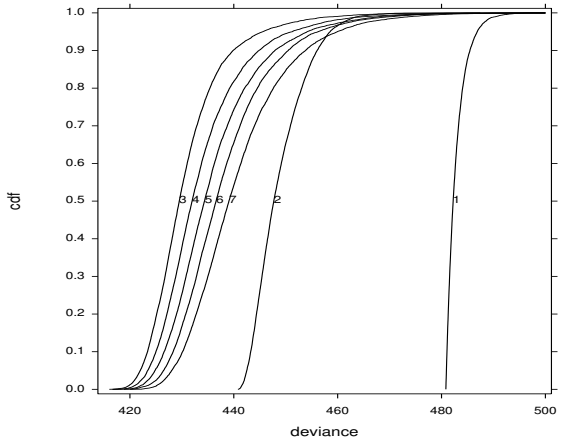


FIGURE 3. Deviances for 1-7 components, galaxy data

6 The current state of statistical inference

The Neyman-Pearson theory is unable to deal with the complexity of current modelling and MCMC analysis. In comparing several competing complex process models for data, the current Bayesian approach based on Bayes factors is unable to provide a reliable comparison. This difficulty is aggravated if the prior hyper-parameters are chosen to "fit the data", to ensure proper convergence of the chain. The resulting integrated likelihoods depend sensitively on these hyper-parameter values, regardless of the sample size.

6.1 A framework for Bayesian model comparisons

The difficulty of Bayes factors with integrated likelihoods can be circumvented by using the full posterior distribution of the likelihood, or deviance, under each model. Given M random draws $\theta_j^{[m]}$ from the converged posterior of the model j parameters θ_j, we substitute them in the model j likelihood $L_j(\theta_j)$ to give M draws $L_j^{[m]} = L_j(\theta_j^{[m]})$ from this likelihood. The likelihoods (or more conveniently deviances) are then compared for *stochastic ordering*; the model with the stochastically largest likelihood distribution (stochastically smallest deviance distribution) is best-supported by the data.

This approach is set out in Aitkin (2010); the galaxy data example from the book is given in Figure 3. The deviance distributions improve dramatically from one to three components, then worsen steadily as K increases beyond three. There is no compelling evidence for more than three components.

References

Aitkin, M. (2001). Likelihood and Bayesian analyses of mixtures. *Statistical Modelling*, **1**, 287-304.

Aitkin, M. (2010). *Statistical Inference: an Integrated Bayesian/Likelihood Approach*. Boca Raton: Chapman and Hall/CRC Press.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.

Jeffreys, H. (1961). *Theory of Probability*. Oxford: Clarendon Press.

Jacobs, R.A., Jordan, M.I., Nolan S.J. and Hinton, D.E. (1994). Adaptive mixtures of local experts. *Neural Computing*, **3**, 79-87.

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

Liu, C.C. and Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, **52**, 362-275.

Pace, L. and Salvan, A. (1990). Best conditional tests for separate families of hypotheses. *Journal of the Royal Statistical Society B* **52**, 125-134.

Postman, M., Huchra, J.P., and Geller, M.J. (1987). Probes of large-scale structures in the Corona Borealis region. *Astronomical Journal*, **92**, 1238-1247.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, **85**, 617-624.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

Silverman, B.W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *Journal of the Royal Statistical Society, Series B*, **47**, 1-53.

Tanner, M.A. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions (3rd edn.)*. New York: Springer.

Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528-540.

Statistical analysis of brains using diffusion tensor images

Ian L. Dryden[1], Alexey Koloydenko[2], Diwei Zhou[3], Li Bai[4]

[1] School of Mathematical Sciences, University of Nottingham, NG7 2RD
[2] Department of Mathematics, Royal Holloway University of London, Egham, Surrey, TW20 0EX
[3] School of Computing and IT, University of Wolverhampton, Wolverhampton, WV1 1LY
[4] University of Nottingham, School of Computer Science, Nottingham, NG8 1BB

Abstract: Statistical analysis of diffusion tensor images is considered, where non-Euclidean distances between tensors are explored. Applications to brain imaging are given, and in particular interpolation and fibre tracking are explored.

Keywords: Covariance; Interpolation; Non-Euclidean; Procrustes; Regularization; Shape.

1 Introduction

The statistical analysis of covariance matrices occurs in many important applications, e.g. in diffusion tensor imaging and longitudinal data analysis. We consider the situation where it is of interest to estimate an average covariance matrix, describe its anisotropy, to interpolate between covariance matrices and to estimate white matter fibre tracks in the brain.
An important difference with standard statistical techniques is that non-Euclidean distances are most natural for comparing covariance matrices, which are symmetric, positive semi-definite matrices.

2 Diffusion tensors

In medical image analysis a particular type of covariance matrix arises in diffusion weighted imaging called a diffusion tensor. The diffusion tensor is a 3×3 covariance matrix which is estimated at each voxel in the brain, and is obtained by fitting a physically-motivated model on measurements from the Fourier transform of the molecule displacement density (Basser et al., 1994).
In the diffusion tensor model the water molecules at a voxel diffuse according to a multivariate normal model centred on the voxel and with covariance

matrix Σ. The displacement of a water molecule $x \in \mathbf{R}^3$ has probability density function

$$f(x) = \frac{1}{(2\pi)^{3/2}|\Sigma|^{1/2}} \exp(-\frac{1}{2}x^T\Sigma^{-1}x).$$

The convention is to call $D = \Sigma/2$ the diffusion tensor, which is a symmetric positive semi-definite matrix. The diffusion tensor is estimated at each voxel in the image from the available MR images. The MR scanner has a set of magnetic field gradients applied at directions $g_1, g_2, \ldots, g_m \in RP^2$ with scanner gradient parameter b, where RP^2 is the real projective space of axial directions (with $g_j \equiv -g_j$, $\|g_j\| = 1$). The data at a voxel consist of signals $(Z_0, Z_1, \ldots, Z_m)$ which are related to the Fourier transform of the displacement density in axial direction $g_j \in RP^2$, $j = 1, \ldots, m$, and the reading Z_0 is obtained with no gradient ($b = 0$). The Fourier transform in axial direction $g \in RP^2$ of the multivariate Gaussian displacement density is given by

$$\mathcal{F}(g) = \int \exp(i\sqrt{b}g^T x)f(x)dx = \exp(-bg^T Dg),$$

and the theoretical diffusion tensor model for the signals is

$$Z_j = Z_0\mathcal{F}(g_j) = Z_0 \exp(-bg_j^T Dg_j), \quad j = 1, \ldots, m.$$

There are a variety of methods available for estimating D from the data $(Z_0, Z_1, \ldots, Z_m)$ at each voxel (see Alexander, 2005), including least squares regression and Bayesian estimation (e.g. Zhou et al., 2008). Noise models include log-Gaussian, Gaussian and Rician noise. A common method for visualizing a diffusion tensor is an ellipsoid with principal axes given by the eigenvectors of D, and lengths of axes proportional to $\sqrt{\lambda_i}$, $i = 1, 2, 3$.
If a sample of diffusion tensors is available we may wish to estimate an average diffusion tensor matrix, investigate the structure of variability in diffusion tensors or interpolate at higher spatial resolution between two or more estimated diffusion tensor matrices.

3 Non-Euclidean statistics

3.1 The Fréchet mean

When comparing diffusion tensors or covariance matrices it is natural to use a non-Euclidean distance $d()$. In this case we must define what is meant by a 'mean covariance matrix'. Consider a probability distribution for a $k \times k$ covariance matrix S on a Riemannian metric space with density $f(S)$. The Fréchet (1948) mean Σ is defined as

$$\Sigma = \arg\inf_{\Sigma} \frac{1}{2}\int d(S, \Sigma)^2 f(S)dS.$$

The Fréchet mean need not be unique in general, although for many distributions it will be (Le, 1995). If we have a sample $S_1, \ldots, S_N$ of i.i.d. observations available then the sample Fréchet mean is calculated by finding

$$\hat{\Sigma} = \arg\inf_{\Sigma} \sum_{i=1}^{N} d(S_i, \Sigma)^2.$$

3.2 Distances between covariance matrices

We now consider specific choices of distances in order to provide estimates of a mean from the sample of N covariance matrices. To ensure the positive definiteness of the covariance matrices, a reparameterization can be used such as $S_i = Q_i Q_i^T$, where $Q_i \in R^{3\times 3}$. For example, $Q_i = chol(S_i)$ is the *Cholesky decomposition*, where Q_i is lower triangular with positive diagonal elements. Note that Q_i and any rotation and reflection of it $Q_i R_i$ ($R_i \in O(3)$) will result in the same S_i, i.e. $S_i = Q_i Q_i^T = Q_i R_i (Q_i R_i)^T, i = 1, ..., N$.
In applications there are several choices of distances between covariance matrices that one could consider, for example see Table 1.

Name	Notation	Form	Estimator
Euclidean	$d_E(S_1, S_2)$	$\|S_1 - S_2\|$	$\hat{\Sigma}_E$
Log-Euclidean	$d_L(S_1, S_2)$	$\|\log(S_1) - \log(S_2)\|$	$\hat{\Sigma}_L$
Riemannian	$d_R(S_1, S_2)$	$\|\log(S_1^{-1/2} S_2 S_1^{-1/2})\|$	$\hat{\Sigma}_R$
Cholesky	$d_C(S_1, S_2)$	$\|Q_1 - Q_2\|$	$\hat{\Sigma}_C$
Root Euclidean	$d_H(S_1, S_2)$	$\|S_1^{1/2} - S_2^{1/2}\|$	$\hat{\Sigma}_H$
Procrustes size & shape	$d_S(S_1, S_2)$	$\inf_R \|Q_1 - Q_2 R\|$	$\hat{\Sigma}_S$
Procrustes shape	$d_F(S_1, S_2)$	$\inf_{R,\beta} \left\| \frac{Q_1}{\|Q_1\|} - \beta Q_2 R \right\|$	$\hat{\Sigma}_F$
Power Euclidean	$d_A(S_1, S_2)$	$\frac{1}{\alpha}\|S_1^\alpha - S_2^\alpha\|$	$\hat{\Sigma}_A$

TABLE 1. Some distances between covariance matrices and notation for the corresponding Fréchet mean estimators, where $Q_i = \text{chol}(S_i), i = 1, 2$.

Estimators $\hat{\Sigma}_E, \hat{\Sigma}_C, \hat{\Sigma}_H, \hat{\Sigma}_L, \hat{\Sigma}_A$ given in Table 1 are straightforward to compute using arithmetic averages. Note that d_S is obtained by optimal rotation/reflection of $chol(S_2)$ onto $chol(S_1)$ using ordinary Procrustes analysis (Dryden and Mardia, 1998). The Procrustes based estimators $\hat{\Sigma}_S, \hat{\Sigma}_F$ involve the use of the Generalized Procrustes Algorithm, which works well in practice (see Dryden et al., 2009). The Riemannian metric estimator $\hat{\Sigma}_R$ uses a gradient descent algorithm which is guaranteed to converge (e.g. see Pennec et al, 2006). In practice it is similar to the log-Euclidean estimator $\hat{\Sigma}_L$ (Arsigny et al., 2007).

We briefly summarize some of the properties of the distances. All distances except d_C are invariant under simultaneous rotation and reflection of S_1 and S_2, i.e. the distances are unchanged by replacing both S_i by $VS_iV^T,\ V \in O(k), i = 1, 2$. Metrics d_L, d_R, d_F are invariant under simultaneous scaling of $S_i, i = 1, 2$, i.e. replacing both S_i by βS_i. Metric d_R is also affine invariant, i.e. the distances are unchanged by replacing both S_i by $AS_iA^T, i = 1, 2$ where A is a general $k \times k$ full rank matrix. Metrics d_L, d_R have the property that $d(A, I_k) = d(A^{-1}, I_k)$, where I_k is the $k \times k$ identity matrix, and d_L, d_R, d_F are not valid for comparing rank deficient covariance matrices. Finally, there are problems with extrapolation with metric d_E: extrapolate too far and the matrices are no longer positive semi-definite (Arsigny et al., 2007).
In some applications covariance matrices are close to being deficient in rank. The Procrustes metrics can deal with certain deficient rank matrices, which is a strong advantage of the approach. A strongly anisotropic diffusion tensor indicates a strong direction of white matter fibre tracts, and plots of measures of anisotropy are very useful to neurologists. A measure that is very commonly used in diffusion tensor imaging is Fractional Anisotropy

$$FA = \left\{ \frac{k}{k-1} \sum_{i=1}^{k} (\lambda_i - \bar{\lambda})^2 / \sum_{i=1}^{k} \lambda_i^2 \right\}^{1/2}, \tag{1}$$

where $0 \leq FA \leq 1$ and λ_i are the eigenvalues of the diffusion tensor matrix. Note that $FA \approx 1$ if $\lambda_1 >> \lambda_i, i > 1$ (very strong principal axis) and $FA = 0$ for isotropy. In diffusion tensor imaging $k = 3$. Other measures include the Procrustes Anisotropy (PA) and Geodesic anisotropy (GA) (see Dryden et al., 2009).

4 Interpolation methods

4.1 Weighted Generalised Procrustes Analysis

Frequently in diffusion tensor imaging it is of interest to interpolate between sets of tensors. The weighted Fréchet sample mean of S_1, ..., S_N at N voxels with a certain distance function $d()$ is defined by:

$$\bar{S} \quad = \quad \arg\inf_{S} \sum_{i=1}^{N} w_i d(S_i, S)^2, \tag{2}$$

where the weights w_i are proportional to a function of the Euclidean distance between voxel locations of the tensors, $0 \leq w_i \leq 1$ and $\sum_{i=1}^{N} w_i = 1$. We choose d_S for the distance and then Weighted Generalized Procrustes analysis (WGPA) is proposed to obtain the weighted mean of S_1, ..., S_N.

The objective of WGPA under rotation and reflection is to minimise a sum of weighted squared Euclidean norms S_{WGPA} which is given by

$$\begin{aligned} S_{WGPA}(S_1,...,S_N) &= \inf_{R_1,...,R_N} \sum_{i=1}^{N} w_i \parallel Q_i R_i - \sum_{j=1}^{n} w_j Q_j R_j \parallel^2 \\ &= \inf_{R_1,...,R_N} \sum_{i=1}^{N} w_i \parallel (1-w_i) Q_i R_i - \sum_{j \neq i} w_j Q_j R_j \parallel^2 \\ &= \inf_{R_1,...,R_N} \sum_{i=1}^{n} \frac{w_i}{(1-w_i)^2} \parallel Q_i R_i - \frac{1}{(1-w_i)} \sum_{j \neq i} w_j Q_j R_j \parallel^2 . \end{aligned} \tag{3}$$

Let $\hat{R}_i, i = 1, ..., N$ be the estimates of the rotation matrices. Then, the *WGPA mean tensor* is given by

$$\bar{S}_{WGPA} = \bar{Q}_{WGPA} \bar{Q}_{WGPA}^T, \tag{4}$$

where $\bar{Q}_{WGPA} = \sum_{i=1}^{N} w_i Q_i \hat{R}_i$. We give Algorithm 1 for estimating $\hat{R}_i, i = 1, ..., N$. Note that the algorithm is guaranteed to converge to a local minimum as the reduction in S_c at each iteration is at least zero.

Algorithm 1 Weighted Generalised Procrustes Method

Initial setting: $Q_i^P \leftarrow chol(D_i)$, $i = 1, ..., N$
S_{WGPA} from previous iteration: $S_p \leftarrow 0$
S_{WGPA} from current iteration: $S_c \leftarrow \sum_{i=1}^{N} w_i \parallel Q_i^P - \sum_{j=1}^{N} w_j Q_j^P \parallel^2$
while $|S_p - S_c| >$ tolerance **do**
  for $i = 1$ to N **do**
    $\bar{Q}_i = \frac{1}{1-w_i} \sum_{j \neq i} w_j Q_j^P$
    Calculate the rotation matrix R_i which minimises $\parallel \bar{Q}_i - Q_i^P R_i \parallel$ with partial ordinary Procrustes analysis
    $Q_i^P \leftarrow Q_i^P R_i$
  end for
  $S_p \leftarrow S_c$
  $S_c \leftarrow \sum_{i=1}^{N} w_i \parallel Q_i^P - \sum_{j=1}^{N} w_j Q_j^P \parallel^2$
end while
$\bar{Q}_{WGPA} \leftarrow \sum_{i=1}^{N} w_i Q_i^P$
return $\bar{Q}_{WGPA}$

4.2 Regularization

In medical image analysis a noisy tensor field may be available and so we wish to carry out regularization. For example, consider a grid of tensors $S_1, \ldots, S_n$ at voxels $x_1, \ldots, x_n$ and we wish to predict the tensor at a new site x. We could use the weighted penalized predictor obtained by minimizing, with respect to Σ,

$$\hat{\Sigma}_{\beta,\omega}(\lambda) = \sum_{i=1}^{n} w_i d(S_i, \Sigma)^\beta + \lambda d_{reg}(\Sigma, \mu)^\omega$$

where the weights $w_i \geq 0, \sum w_i = 1$ are functions of the distance to the new site, $\lambda > 0$ is a regularization parameter, and μ is a reference matrix, such as the identity matrix, zero matrix or an overall average. For example we could use $w_i \propto \exp\{-\gamma \|x - x_i\|^2\}, i = 1, \ldots, n$. Note that the choice of distances for d and d_{reg} need not be the same.
Consider now smoothing across an image at the voxels $x_1, \ldots, x_n$, and so we need to minimize, with respect to $\Sigma_j, j = 1, \ldots, n$,

$$\sum_{j=1}^{n} \sum_{i=1}^{n} w_{ij} d(S_i, \Sigma_j)^\beta + \lambda \sum_{j=1}^{n} d_{reg}(\Sigma_j, \mu)^\omega,$$

and w_{ij} is the weight as a function of the distance between sites i and j. Note $(\beta, \omega) = (2, 0)$ gives the weighted Fréchet mean, if $(\beta, \omega) = (\beta, 0)$ we have a type of M-estimator (Kent, 1992; Dryden and Mardia, 1998, p298), if $(\beta, \omega) = (1, 0)$ we have the geometric median (Fletcher and Joshi, 2009), if $(\beta, \omega) = (2, 2)$ non-Euclidean type of ridge-regression, and if $(\beta, \omega) = (2, 1)$ a non-Euclidean type of LASSO (see Tibshirani, 1996). Note that for the power metric (and Euclidean and square root) the space is Euclidean, and so using this procedure is relatively straightforward in this case.

5 Applications

5.1 Interpolation

A tensor field from a healthy human brain has been smoothed and interpolated (with 2 interpolations between each pair of original voxels). The Fractional Anisotropy (FA) maps from the processed tensors are shown in Figure 1. Obviously, the FA map from the processed tensor data is much smoother than the one without processing. The feature that the cingulum is distinct from the corpus callosum is clearer in the anisotropy map from the processed data than those without processing in Figure 1.

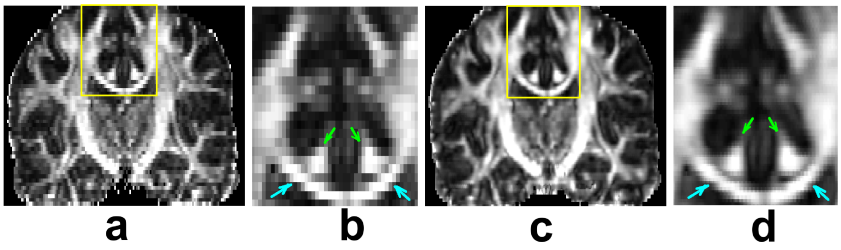


FIGURE 1. Smoothing and interpolation of the diffusion tensor data from human brain. a: FA map from Bayesian tensor field. c: FA map from processed tensor field. b and d: Zoomed inset regions. Arrows point down to the cingulum and point up to the corpus callosum.

5.2 Tractography

As a final application we give some initial results of fibre tractographies of the brain stem of a healthy human in Figure 2. It is of great interest to study the white matter fibre tracts in the brain in order to explore connectivity between different parts, both in healthy and patient brains. From different seed points in the brain stem, white matter fibres are tracked by following interpolated paths of principal directions from diffusion tensors. Tractography from the WGPA processed tensor field is different from the other methods, and work is currently underway to assess whether WGPA is preferable.

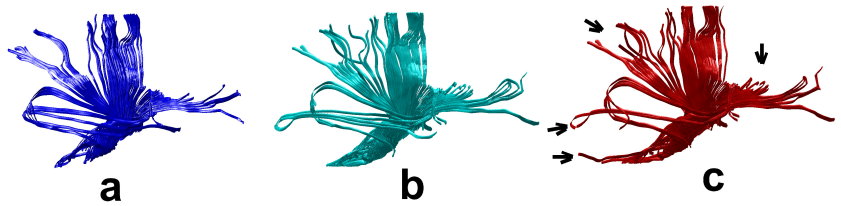


FIGURE 2. Fibre tractograhpies using the Bayesian estimates (a), Euclidean smoothing (b) and WGPA smoothing (c). Black arrows point out some obvious differences of the WGPA tracts compared with other methods.

6 Conclusions

Methodology for estimation and inference in the space of covariance matrices has application in many areas, including diffusion tensor imaging, structural tensor analysis in computer vision, and modelling longitudinal data with Bayesian and random effect models. There are many choices of metric available, each with its advantages. The particular choice of what is best will depend on the particular application. The use of the Procrustes

size-and-shape metric d_S is particularly appropriate when the covariance matrices are close to being deficient in rank.

References

Alexander, D. C. (2005). Multiple-fiber reconstruction algorithms for diffusion MRI. *Ann NY Acad Sci*, 1064:113–133.

Arsigny, V., Fillard, P., Pennec, A. and Ayache, N. (2007). Geometric Means in a Novel Vector Space Structure on Symmetric Positive-Definite Matrices, *SIAM Journal on Matrix Analysis and Applications*, **29**, 328–347.

Basser, P. J., Mattiello, J., and Le Bihan, D. (1994). Estimation of the effective self-diffusion tensor from the NMR spin echo. *J Magn Reson B.*, **103**, 247–254.

Dryden, I. L., Koloydenko, A., and Zhou, D. (2009). Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Annals of Applied Statistics*. **3**, 11021123.

Dryden, I. L. and Mardia, K. V. (1998). *Statistical Shape Analysis*. Wiley, Chichester.

Fletcher, P.T., Venkatasubramanian, S. and Joshi, S. (2009). The geometric median on Riemannian manifolds with application to robust atlas estimation. 45, S143-S152.

Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H. Poincaré*, 10:215–310.

Kent, J. T. (1992). New directions in shape analysis. In Mardia, K. V., editor, *The Art of Statistical Science*, pages 115–127. Wiley, Chichester.

Le, H.-L. (1995). Mean size-and-shapes and mean shapes: a geometric point of view. *Advances in Applied Probability*, 27:44–55.

Pennec, X., Fillard, P. and Ayache, N. (2006). A Riemannian Framework for Tensor Computing, *Int. J. Comput. Vision*, **66**, 41–66.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, Vol. 58, 26-7-288).

Zhou, D., Dryden, I. L., Koloydenko, A., and Bai, L. (2008). A Bayesian method with reparameterisation for diffusion tensor imaging. In Reinhardt, J. M. and Pluim, J. P. W., editors, *Proceedings, SPIE conference. Medical Imaging 2008: Image Processing*, page 69142J.

Bayesian computing with INLA

Håvard Rue[1]

[1] Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway

Abstract: In this talk I will discuss how to do Bayesian inference in the class of latent Gaussian models (LGM). LGM is perhaps the most widely used model construct in statistics, and, perhaps surprisingly, all marginals can be very well and quickly approximated without the need for simulation based inference. Further, it seems like in most cases, the approximated marginals are more accurate than the ones found from long MCMC runs.

Keywords: Bayesian computing; Latent Gaussian Models; Gaussian Markov Random Fields; Sparse Matrices

1 Introduction

In the core of Bayesian statistics is the computational machinery to compute posterior marginals. In the "old days", Laplace approximations and numerical integration were the main tools (Tierney and Kadane, 1986). After the introduction of MCMC, these tools just "died out" and MCMC quickly became the main tool. In this talk, I'll return to Laplace approximations and numerical integration, and demonstrate that for a particular class of models, *latent Gaussian models* (Rue and Held, 2005), then these tools can provide superior approximations to posterior marginals over MCMC (Rue, Martino and Chopin, 2009). There is of course something with small print; we have to apply various tricks and numerical methods for sparse matrices in order to achive a practical tool. However, when this tool *is available*, then it appearently solves the Bayesian computational problem in many of the most used models.

1.1 Latent Gaussian Models

I will now introduce the class of latent Gaussian models (LGM) before looking at some examples.
By a *latent Gaussian models* we mean the following three stages hierarchical model

Stage 1 The data $y = (y_i)$ are described conditionally on some latent random variables $x = (x_i)$ and possiby some unknown (non-Gaussian)

variables θ_1, as

$$\pi(y \mid x, \theta_1) = \prod_{i \in \mathcal{I}} \pi(y_i \mid x_i, \theta_1)$$

over some subset of the indices $\mathcal{I} \subset \{1, \ldots, n\}$.

Stage 2 The latent process x is jointly Gaussian conditionally on some unknown parameters θ_2,

$$x \mid \theta_2 \ \sim \ \mathcal{N}(\mu, Q(\theta_2)^{-1}).$$

Here, μ is fixed and the precision matrix Q can depend on θ_2.

Stage 3 The (typical) non-Gaussian variables $\theta = (\theta_1, \theta_2)$ have some prior $\pi(\theta)$.

The surprising fact is that many of the applied statistical models are of this type; for example most but not all of models in dynamic linear models, stochastic volatility models, generalised linear (mixed) models, generalised additive (mixed) models, spline smoothing, semiparametric regression, space-varying regression models, disease mapping, log-Gaussian Cox-processes, model-based geostatistics, spatio-temporal models and so on. Note that the notion of LGM is not meant to help or aid the modeling problem, but only to unify their inference.
As an example, consider the following simple "mixed effect" model, for $i = 1, \ldots, n$, $j = 1, \ldots, n_i$ and fixed covariates z,

$$\begin{aligned}
\eta_{ij} &= a + bz_i + u_i + v_{ij} \\
u_i &\overset{iid}{\sim} \mathcal{N}(0, \tau^{-1}) \\
v_{ij} &\overset{iid}{\sim} \mathcal{N}(0, \kappa^{-1}) \\
a &\sim \mathcal{N}(0, \kappa_a^{-1}) \\
b &\sim \mathcal{N}(0, \kappa_b^{-1}) \\
y_{ij} | \eta_{ij} &\sim \text{Poisson}(\exp(\eta_{ij})) \\
\tau &\sim \text{Gamma}(1, 0.00005) \\
\kappa &\sim \text{Gamma}(1, 0.00005).
\end{aligned}$$

Due to the additive form in the linear predictor,

$$x = (\eta, a, b, u, v) \ \sim \ \mathcal{N}(0, Q(\theta_2)^{-1})$$

for some (singular) precision matrix Q and (hyper-)parameter $\theta_2 = (\tau, \kappa)$, and θ_1 is the empty set.
A charaterisitic feature of LGM, is that the dimension of the latent field, x, is often large, say $10^2 - 10^5$, while the dimension of θ is small. In the previous example, $\dim(\theta)$ is 2, while $\dim(x) = 2\sum_i n_i + 2 + n$.

2 The Laplace Approximation

The classical Laplace approximation, from a statistical point of view, is an technique to integrate out random variables. In the context of LGM, we can write

$$\pi(\theta \mid y) \propto \frac{\pi(\theta)\pi(x|\theta)\prod_{i\in\mathcal{I}}\pi(y_i \mid x_i,\theta)}{\pi(x|y,\theta)}$$

dropping the subscript for the hyperparameters. The numerator is known but the full conditional for x is not known for non-Gaussian data. However, this does not mean that we cannot approximate the full conditional for x. The Laplace approximation is to approximate the full conditional of x with a Gaussian approximation indexed by θ (matching the mode and curvature around the mode), and evaluating the right hand side at the mode (which is a function of θ). In the case of LGM, we can expect that the Gaussian approximation is quite good, as

1. without or with Gaussian data, the Gaussian approximation to the full conditional for x is exact,

2. for non-Gaussian data, the the error in the approximation comes from 3rd and higher order terms of the log-likelihood (seen as a function of x_i), while the Gaussian prior for x will reduce the error by "borrowing strength" within x.

Asymptotic results about the Laplace approximation say that, for replicated observations, then the normalised approximation of $\pi(\theta|y)$ has an *relative* error of $\mathcal{O}(N^{-3/2})$ where N is proportional to the number of observations. This results holds even if $\pi(x|\theta)$ is non-Gaussian, but we expect the error constant to be much smaller in the Gaussian case.
An important fact is that the error is *relative* and not additive as for Monte Carlo methods. This means that we are also able to estimate the tail of $\pi(\theta|y)$ accurately.
A complicated fact is that, for most LGM, the natural asymptotic schemes do not involve replicated observations, but schemes where $\dim(x) = \mathcal{O}(\dim(y))$. Although there are some theoretical results available, so we need to verify results empirically as well; see Rue, Martino and Chopin (2009) for a detailed discussion.

3 The Integrated Nested Laplace Approximation

We will assume that the main goal of the inference is to compute/approximate the posterior marginal for x_i and θ_j,

$$\pi(\theta_j \mid y) = \int \pi(\theta \mid y)\, d\theta_{-j}, \quad \text{and} \quad \pi(x_i \mid y) = \int \pi(\theta \mid y)\, \pi(x_i \mid y,\theta)\, d\theta,$$

for all i and j. Here θ_j is the jth element of the θ-vector. Letting $\widetilde{\pi}(\cdot)$ denote an approximation, it is clear that we can express the approximated posterior marginals as

$$\widetilde{\pi}(\theta_j \mid y) = \int \widetilde{\pi}(\theta \mid y)\, d\theta_{-j}, \quad \text{and} \quad \widetilde{\pi}(x_i \mid y) = \int \widetilde{\pi}(\theta \mid y)\, \widetilde{\pi}(x_i \mid y, \theta)\, d\theta,$$

This representation takes advantage of the structure of the LGMs:

1. The dimension of θ is small hence numerical integration of $\widetilde{\pi}(\theta|y)$ is feasible. Also, to compute $\widetilde{\pi}(\theta|y)$ we need to approximate $\pi(x|\theta, y)$ which is *approximate* Gaussian.

2. A similar comment applies to $\widetilde{\pi}(x_i|y, \theta)$, for which we must approximate $\pi(x_{-i}|x_i, \theta, y)$.

4 Doing the work

Although we have sketched the overall scheme for computing the marginals, it is at this point not indeed clear that this will "work" in practice, as the dimension of x is potentially larger and somebody has to do the work!
The simplifying factor is that, for (most) LGM, the Gaussian approxmation to $\pi(x|y, \theta)$ will be a *Gaussian Markov random field* (GMRF). GMRFs are an easy construct: they are jointly Gaussian distributed with additional conditional independence properties,

$$x_i \perp x_j \mid x_{-ij}$$

for some pairs $i \neq j$. The nice feature is that conditional independence of x_i and x_j is equivalent to that particular element of the precision matrix, Q_{ij}, being zero. Hence, a lot of conditional independence is equivalent to a (very) sparse precision matrix. Precision matrices can be factorised much quicker than dense matrices, and for typical LGM in time and space, the cost will be $\mathcal{O}(n)$ and $\mathcal{O}(n^{3/2})$, respectively, compared to $\mathcal{O}(n^3)$ for dense matrices.
The simplest example of a GMRF, is an auto-regressive process (in time), which is often expressed in an imprecise notation as

$$x_t = \phi x_{t-1} + \epsilon_t, \qquad t = 1, \ldots, n, \quad |\phi| < 1,$$

with precision matrix H, say. In this case, H is tridiagonal and only $3n-2$ of its n^2 elements are non-zero. Note that the non-zero structure of H does not depend on the actual numerical value of ϕ which controls the correlation

$$\text{corr}(x_t, x_{t+h}) = \phi^{|h|}.$$

5 INLA in practice

It should not be a surprise at this stage that there are many missing details of how these approximations actually are computed, and that "tricks" are required for fast computation. For some details about the most imporant ideas, I refer to the main reference (Rue, Martino and Chopin, 2009).
To illustrate its use, I will demonstrate it on a simple simulated example. To get started, you will need to install the INLA-library. It is not resident at CRAN (for security issues) since it contains prebuilt binaries of the inla-program; the program that does the work.
Unless you already have the INLA-library installed, do

```
> source("http://www.math.ntnu.no/~hrue/givemeINLA.R")
```

which will istall the library. To load it, then do

```
> library("INLA")
```

and we are ready to go. To make it simple, we will simulate the "mixed"-model previously described using the following code (use default priors only)

```
> set.seed(123)
> a = 0
> b = 1
> n = 25
> tau = 1
> u = rnorm(n, sd = sqrt(1/tau))
> ni = sample(1:n, size=n, replace=TRUE)
> z  = rnorm(n)
> sum.ni = sum(ni)
> kappa = 2
> v = rnorm(sum.ni, sd = sqrt(1/kappa))

> eta = numeric(sum.ni)
> yy = numeric(sum.ni)
> ii = numeric(sum.ni)
> jj = numeric(sum.ni)
> zz = numeric(sum.ni)
> k = 1
> for(i in 1:n) {
>     for(j in 1:ni[i]) {
>         eta[k] = a + b*z[i] + u[i] + v[k]
>         yy[k] = rpois(1, lambda = exp(eta[k]))
>         ii[k] = i
>         jj[k] = k
>         zz[k] = z[i]
>         k = k+1
>     }
> }
```

To do the Bayesian analysis, we specify the `formula` and call `inla`

```
> formula = yy ~ 1 + zz + f(ii, model="iid") +
                  f(jj, model="iid")
> result = inla(formula, family = "poisson",
                data = data.frame(yy,ii,jj,zz))
```

Output from the analysis is most easily viewed with

```
> summary(result)        or    > plot(result)
```

A edited summary of the results are as follows

```
Fixed effects:
                     mean           sd    0.025quant ...
(Intercept) 0.06368792089 0.2089212260 -0.3633878831 ...
zz          1.06807240228 0.2178344761  0.6476383295 ...

Model hyperparameters:
                 mean   sd     0.025quant 0.5quant 0.975quant
Precision for ii 1.2668 0.4485 0.5945     1.1993   2.3341
Precision for jj 2.9835 0.6296 1.9431     2.9149   4.4
```

Note that properties like the mean, sd and the quantiles, are derived from the approximated marginals.
The Markov structure in the latent field x is evident from Figure 1, which displays the precision matrix, the reordered precision matrix and its Cholesky factorisation. In total, the analysis took less than than one second on my laptop.

6 Summary

In this extended abstract I have briefly introduced the ideas behind INLA and its use for doing approximate Bayesian inference for latent Gaussian models. However, this illustration does not convey the wide range of examples, in various fields, which can be expressed as a LGM and where INLA can be applied. The www-page

```
www.r-inla.org
```

contains all the software, source code, and various worked examples; please visit!

References

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, 71(2):319–392.

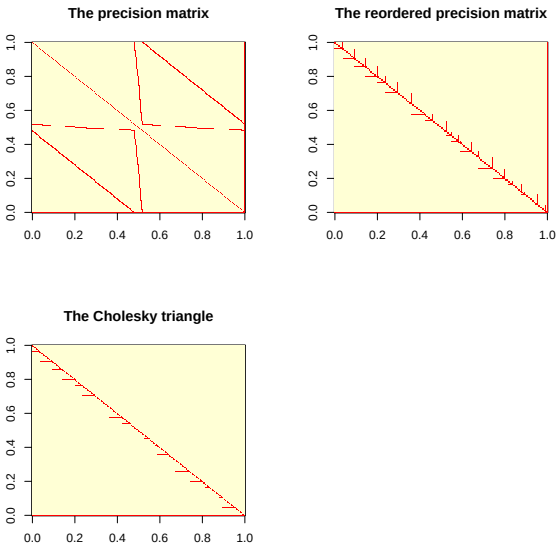


FIGURE 1. The precision matrix, the reordered precision matrix and its Cholesky triangle for the example.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86.

A statistical framework for an evidence base to support environmental regulation and policy.

E. Marian Scott[1], A. Bowman[1], C. Ferguson[1], D. Lee[1], D. O'Donnell[1], M. Franco Villoria[1], J. Campbell Gemmell[2]

[1] Dept. of Statistics, University of Glasgow, Glasgow, G12 8QW
[2] Scottish Environment Protection Agency, Stirling, FK9 4TR

1 Introduction

A recent article in the Economist, Feb 2010, stated "Chief information officers (CIOs) have become somewhat more prominent in the executive suite, and a new kind of professional has emerged, the data scientist, who combines the skills of software programmer, statistician and storyteller/artist to extract the nuggets of gold hidden under mountains of data. Hal Varian, Google's chief economist, predicts that the job of statistician will become the "sexiest" around. Data, he explains, are widely available; what is scarce is the ability to extract wisdom from them."
When we consider the environment, with its associated statements of policy and regulation, there are similarly issues and situations where "the wisdom is not yet extracted" or where we are "*data rich and information poor*". There have been many statements of a similar nature to "How much or how little we know about the links between environmental policy measures and their actual impact in the environment", or "much of the information gathered is of limited use in assessing the impact of environmental measures" (Nigel Haigh, foreword of Environmental Issues, Report 25/EC).

2 State of the environment reporting

Evaluation of the effectiveness of policy requires reporting on the state of the environment (SoE), and such reports are key parts of the evidence base for politicians, the regulators and the public (CAMERAS, SG, 2009). In terms of reporting on the effectiveness of a policy or a potential issue, there are several questions to consider:
Question 1: What has changed and in what way?
Question 2: Can we attribute those changes to any actions?
Question 3: Are the changes significant?

It should be self-evident that Statistics and statisticians have an important role to play in reporting since certainly the first question lies within the remit of the statistician, and questions 2 and 3 require a multi-disciplinary approach, including Statistics.

3 Statistical and Environmental Challenges

3.1 Environmental monitoring and sampling

Environmental policy documents are typically phrased in terms of objectives, targets, guide values, standards and reductions relative to a baseline and cover all aspects of our environment including water, air and noise pollution; waste management, harmful substances, radioactivity, protection of wildlife and countryside, as well as global issues such as climate change. Many policy documents also stipulate in some form how data, in support of the policy effectiveness, need to be collected. As an example, the Water Framework Directive (WFD 2000/60/EC) states: "Member states must ensure that enough individual water bodies of each water type are monitored and determine how many stations are required to determine the ecological and chemical status of the water body".
Routine monitoring data serve many purposes (and are often used for a purpose for which they were not designed), including:

- assessment of long-term changes in natural conditions
- quantifying any impacts of accidental pollution
- or new developments assessing compliance with directives

In Scotland, the environment protection agency (SEPA), for purposes of the water framework directive describes their strategy as risk-driven. Further in terms of their monitoring network, they classify sites in three ways, **surveillance**, **operational** and **investigative**.
A **surveillance** network is spatially distributed, and used to assess long-term changes in natural conditions and long-term changes due to widespread anthropogenic activity, while **operational** monitoring is driven by risk assessments and located in areas of known risk. **Investigative** monitoring is a more variable network responsive to unplanned events and emerging risks, where the source of the risk (the pressure), is not always well understood. "Investigative monitoring should be carried out where the reason for any exceedances is unknown, or where surveillance monitoring indicates that the objectives set out are not likely to be achieved and operational monitoring has not already been established, in order to ascertain the causes of a water body or water bodies failing to achieve the environmental objectives". (SEPA, 2005)

From a statistical standpoint, one challenge in using routine monitoring data may come from the recognition that sampling locations are often not selected to be representative, but are preferentially chosen (Diggle et al, 2010) with implications for subsequent statistical modelling.

3.2 Environmental indicators

The main form of official environmental data processing and information presentation comes in the form of indicators. Environmental indicators are powerful tools that serve many purposes. They serve both as evaluative tools but also as a means of sharing information, especially with the public, and are especially used to illustrate trends over time and so are directly relevant to Question 1.

"*environmental indicators have a crucial role to play in the simplification, quantification, standardisation and communication of environmental conditions to regulators and policy- makers*" (Johnson, ICES, 2008)

According to OECD (2003), criteria for indicator selection are "policy relevance and utility for users, analytical soundness and measurability". There are several definitions of an indicator, e.g. OECD states "it is a parameter or value derived from a parameter that provides information about the state of the environment, and has a significance extending beyond that directly associated with any given parametric value". There are also extensions to create a composite index, e.g. OECD (2003) state that "an environmental index is a set of aggregated or weighted parameters or indicators", therefore it is a composite (multi-variable) numerical value related to state of the environment (e.g taking an ecosystem approach), individual indicators are compiled (typically as a weighted average) into a single index. Composite indicators could therefore be used to measure such multidimensional concepts, as for example sustainability. Frequently (almost always) there is no assessment of uncertainty. The data on which the indicator is based are often not representative (and sparse in space), so it is hard to build a spatial aggregation model to identify what is happening (Question 1). If the indicator is highly aggregated, then it may indeed be difficult to disentangle the effects of pressures such as climate change (Question 2).

3.3 Statistical challenges

When confronted with policy needs, the nature of regulation, the relatively restricted nature of the routine data and the complexity of some of the environmental issues, there are a number of statistical, computational, and modelling challenges. Definition and detection of trends and the assessment of changes, including abrupt change is perhaps the most common challenge, but when set in a spatial context, spatio-temporal modelling, and describing the spatial correlation structure make this problem a statistical research issue. Other common problems with a clear statistical interest include:

creation of composite indicators; dealing with high (or low) quantiles of a distribution (evaluating risks) and modelling extremes (in space and time); describing relationships; and handling and quantifying uncertainties.

4 Case studies

In this section, four case studies are briefly reviewed in the light of the foregoing discussion, regarding policy and effectiveness evaluation. Case studies 1 and 2 relate to the EU Water Framework Directive, while Case study 3 deals with the Flooding directive and case study 4 looks at air quality. Each case study illustrates some distinctive statistical challenges.

4.1 Case study 1: Water quality in Loch Leven

(joint work with C. Ferguson, A. Bowman and L. Carvalho (CEH))

Loch Leven (Kinross, Scotland) is the largest shallow lake in Great Britain and the Centre for Ecology & Hydrology (CEH), in Edinburgh, has monitored the loch since 1967. Over this period the loch has experienced eutrophication (a deterioration in water quality as a result of increased algae from increased nutrient loading) and subsequent recovery, following reductions in nutrient loading from the catchment. The time series also spans a period of changing climate, with increasing temperatures in winter and spring; details are provided in Carvalho & Kirika (2003). This makes Loch Leven an ideal site for investigating statistical techniques that can differentiate between these features. The measurement series covers more than 30 years, including a variety of biological, chemical and hydrological indicators but with changes in sampling frequency during this period. One of the key questions of interest is how climate change is affecting the loch (Questions 1 and 2)? Statistical modelling using additive and varying coefficient models to look at trends and effects of climate (water temperature) on some of the key water parameters have been used, one example of which is shown below.

An additive model was fit for chlorophylla (phytoplankton) of the form

$$\begin{aligned} \log(\text{chlorophylla})_t &= \alpha + m_1(\text{year}_t) + m_2(\text{month}_t) + m_3(\log(\text{SRP})_t) \\ &\quad + m_4(\log(\textit{Daphnia})_t) + m_5(\text{watertemp}_t) \\ &\quad + m_6(\log(\text{NO}_3 - \text{N})_t) + \epsilon_t \end{aligned}$$

where SRP, NO3-N are key nutrients, and Daphnia are zooplankton (water fleas). In the model, the errors are assumed to have variance matrix $V\sigma^2$ where V is a correlation matrix based on an AR(1) process and since the month term in each of these models is defined on a cyclical scale, a circular weight function was used for this component (Ferguson et al, 2007).

4.2 Case study 2: Spatial patterns of change in river basins

(joint work with A. Bowman, D. O'Donnell and M. Hallard (SEPA))

As part of the Water Framework Directive, river basin management plans should be drawn up to cover all the rivers, lochs/lakes, estuaries, coastal waters, groundwater and artificial waters (such as reservoirs) in a river basin district. This case study refers to a statistical analysis of the river monitoring network on the River Tweed, which is in the south west of Scotland. Data are available from January 1986 to October 2006 for up to eighty five monitoring sites on the river, with the number of actual sites varying over time. A number of different chemical and biological determinands are measured, but here the focus is on diffuse pollution as a result of sewage effluent and runoff from fertilisers used on agricultural land. Spatial patterns of change may be important and interpolation over the entire network based on the monitoring stations is possible, but needs a spatial model which captures the specific river network. Until recently, a Euclidean distance based model would have been commonly used, but recent developments have seen the creation of a river distance model (where river distance is defined as the shortest distance between two locations, along the river network), or a mixed model which has both river and Euclidean based distance measures. (Cressie et al 2006, ver Hoef et al, 2006)
It is possible to define a class of covariance models which are suitable for use with stream distance instead of Euclidean distance. Ver Hoef et al. (2006) state that these models take the form shown below where $h_{str}(s_i; t_j)$ is the stream distance between points s_i and t_j . The model shown below not only uses stream distance but it also incorporates two other potentially desirable features: a weighting structure which can be used to weight according to how large the stream is (in terms of some measure such as flow or stream order) and zero covariance assigned to any pair of stations that are flow-unconnected. The covariance model $C_1(h)$ can be chosen from the standard set of autocovariance functions, such as the exponential, spherical, linear with sill etc. This means that, by using a weighing structure along with existing models, a class of valid autocovariance models that use stream instead of Euclidean distance can be created.

$$\begin{aligned} C(s_i; t_j|\theta_j) &= 0 && \text{if } s_i \text{ and } t_j \text{ are not flow connected} \\ &= C_1(0) + \nu_j^2 && \text{if } s_i = t_j \\ &= \prod_k \sqrt{w_k} C_1(h_{str}) && \text{otherwise} \end{aligned}$$

4.3 Case study 3: Flood risk in Scotland

(joint work with Maria Franco Villoria and Trevor Hoey)

The Flood Risk Management (Scotland) Act 2009 was enacted on June 16, 2009 introducing new approaches to flood risk management and which is suited to the impact of climate change. Of specific interest is the assessment of flood risk, and the spatial aspects of extremes in river flow. River flow records have formed the basis of many flood risk estimates, their modeling based on classical statistical models, that have assumed stationarity. However, under climate change and climate change scenarios, there is an expectation that the flow series may no longer be stationary, therefore statistical models that do not make this assumption are required. In Scotland, these changes may be apparent in west to east differences in terms of rainfall and river flow. Wavelet analyses are being applied to individual river flow series to identify the local behaviour of potentially non stationary series, resulting in a time-frequency representation of the data (Percival and Walden, 2006). The resulting representations of variability can then be explored in terms of climatic drivers (such as the North Atlantic Oscillation and Atlantic Meridional Oscillation), and to explore spatial differences.
However, the potentially more interesting statistical question concerns the spatial pattern in extreme flows which is one of considerable scientific interest in the wider context of climate change. Generalisation and extension from the univariate extreme value theory is complex, with recent developments exploring max-stable stochastic processes (Schlather, 2002), and building spatial hierarchical models (Sand and Gelfand, 2009).
Thirty five rivers of different catchment sizes across Scotland were selected on the basis of data quality and quantity, and the monthly maxima were calculated for the time period January 1985 to December 2005. In a preliminary analysis, at each site, the monthly overall mean was removed to de-seasonalise the series and then, a GEV distribution was fitted separately to each of 35 series and a geostatistical analysis of the parameter estimates carried out.
Approaches like these will allow us to address, within the context of the legislation, what the flood risks are (using the historical data), to evaluate complex change in time, and to explore the drivers of change.

4.4 Case study 4: Air quality indicators

(joint work with Duncan Lee and Claire Ferguson)

Air pollution levels are routinely monitored at both urban, rural and background sites to determine whether targets set as part of the air quality strategy (DEFRA, 2007) are met. The commonly monitored pollutants include carbon monoxide, ozone, nitrogen dioxide and particulate matter (PM10), which all may have potential human health implications. As a result of policy and strategy needs, indicators have become the instrument

of choice for reporting on air quality, typically constructed first as an aggregation over space (and possibly time) before then being averaged over pollutants (Bruno and Cocchi, 2002). Simplistic aggregations may have a number of undesirable properties, or unresolved issues, including using potentially quite sparse (and almost certainly non-spatially representative) monitoring sites to create a spatial average, and the issue of weighting the different pollutants. Recent work has included developing a Bayesian geostatistical model (delivering uncertainty ranges for the spatial aggregation, that can also be extended to deal with preferential sampling issues). (Lee et al, 2010). Statistical questions concern the construction, properties and uncertainties associated with such indicators, as well as design issue for monitoring networks.

5 Conclusions

Sophisticated statistical models for trends can give added value to routine monitoring data, provide better descriptions of complex change behaviour, including creation of indicators and indices with desirable properties and begin to tease out climate change driven effects in environmental quality, but as a community we need to be convincing in these benefits. To deliver the information (contextual and interpretational) value of routine monitoring data, greater innovative statistical analysis is needed. Uncertainty-evaluation and representation is vital but often missing and we could ask why this should be. There is often unease about presenting uncertainty.
Many environmental issues are set within a spatio-temporal framework and there are challenges in how to build a spatial correlation model that respects the natural system, which for spatio-temporal models, presents an even more complex problem, compounded by the fact that the spatial locations may be sparse and their selection process frequently spatially non-representative. In the policy context, indicators remain the communication tool of choice, but they too present statistical challenges such as how to define an indicator for an ecosystem assessment, and how to define weights to give a valid and robust composite index, such as for sustainability. As statisticians and citizens, we have a challenge!

References

Bruno, F., Cocchi, D. (2002). A unified strategy for building simple air quality indices. *Environmetrics*, **13**, 243261.

CAMERAS. Scottish Government web site, downloaded April 2010.

Carvalho, L., and Kirika, A. (2003). Changes in shallow lake functioning: response to climate change and nutrient reduction. *Hydrobiologia* **506-509**, 789-796.

Cressie, N., J. Frey, B. Harch, and M. Smith (2006). Spatial prediction on a river network. *Journal of Agricultural, Biological, and Environmental Statistics* **11 (2)**, 127-150.

Department for Environment, Food and Rural Affairs (2007). The Air Quality Strategy for England, Scotland, Wales and Northern Ireland (Volume 1). Her Majesty's Stationary Office.

Diggle, P. J., R. Menezes, and T. Su (2010). Geostatistical inference under preferential sampling. JRSS(C), 59(2), 191-233.

Economist Feb 2010. Data, data everywhere.

Environmental Issues, Report 25/EC. European Environment Agency, Copenhagan.

Ferguson C A, Bowman A W, Scott E M, Carvalho L (2007). Model Comparison for a complex ecological system. *JRSS (A)*, **170 (3)** 691-711.

Johnson M (2008). Environmental indicators: their utility in meeting the OSPAR Convention's regulatory needs. *ICES J of Marine Science* **65(8)**, 1387-1391

Lee D, Ferguson C A, Scott E M (2010). Constructing representative air quality indicators with measures of uncertainty. JRSS(A) to appear.

OECD (2003). OECD Environmental Indicators. Development, measurement and use. Reference paper. Downloaded April 2010,

Percival, D.B. and Walden, A.T. (2006). *Wavelet Methods for Time Series Analysis*. Cambridge University Press, Cambridge Series in Statistical and Probabilistic Mathematics.

Sang, H., Gelfand, A.E. (2009). Hierarchical modeling for extreme values observed over space and time. Environ Ecol Stat. In press.

Schlather M (2002). Models for stationary max-stable random fields. *Extremes* **5(1)**, 33-44.

SEPA (2005). RSG(05)07 & 09. WFD Environmental Monitoring Network Development Downloaded April 2010.

Water Framework Directive (2000). "Directive 2000/60/EC of the European Parliament and of the Council establishing a framework for the Community action in the field of water policy" downloaded April 2010 from EC/Environment web site.

Ver Hoef, J. M., E. Peterson, and D. Theobald (2006). Spatial statistical models that use flow and stream distance. *Environmental and Ecological Statistics* **13 (4)**, 449-464.

Longitudinal Structural Mixed Models for the Analysis of Surgical Trials with Noncompliance

Colleen M Sitlani[1], Patrick J Heagerty[1], Emily Blood[2], Tor Tosteson[2]

[1] Department of Biostatistics, University of Washington, Seattle, WA
[2] Dartmouth College, Hanover, NH

Abstract: Patient noncompliance complicates the analysis of many randomized surgical trials that seek to evaluate the causal effect of surgical intervention as compared to a non-surgical treatment. If selection for treatment depends on intermediate patient characteristics or outcomes, then "as-treated" analyses may be biased for the estimation of causal effects. Therefore the selection mechanism for treatment and/or compliance should be carefully considered when estimating the causal effect of intervention. We compare the performance of analysis methods when endogenous processes lead to patient crossover in the presence of an underlying longitudinal structural mixed model that is a natural example of a structural nested mean model. Standard linear mixed models will be valid under selection mechanisms that depend only on past covariate and outcome history. If there are underlying patient characteristics that influence selection, then these likelihood methods can be extended via maximization of the joint likelihood of exposure and outcomes. Causal estimation methods such as marginal structural models, g-estimation and instrumental variable approaches can also be valid and we review their implementation in this setting. The assumptions required for valid estimation vary across these approaches; thus the choice of methods for analysis should be driven by which selection assumptions are plausible.

1 Introduction

Randomized trials with the goal of evaluating the long-term benefit of a surgical intervention as compared to non-surgical treatment are often faced with serious patient noncompliance. Frequently subjects assigned to surgery delay or subsequently refuse surgery, while non-surgical subjects may ultimately seek and receive surgery. For example, in the Spine Patient Outcomes Research Trial (SPORT) evaluation of surgery for lumbar invertebral disk herniation, 60% of those randomized to surgery actually received it, while 45% of those randomized to nonoperative treatment subsequently received surgery (Weinstein *et al.*, 2006). Standard intent-to-treat (ITT) analyses do not capture the full efficacy of surgery in the presence of such noncompliance (Flum, 2006; Ellenberg, 1996). In an effort to estimate

the efficacy of treatment, investigators often supplement ITT analyses with "as-treated" analyses that incorporate subjects' actual treatment received instead of their assigned treatment. There are several statistical challenges associated with longitudinal 'as-treated' analyses which seek to estimate average causal effects (ACEs) attributable to surgery (Flum, 2006; Ellenberg, 1996).

The goal of this manuscript is to overview both statistical models and estimation methods that may be appropriate for the analysis of surgical non-compliance. We first anchor the statistical objective by defining the causal effect of interest based on a structural model for the longitudinal outcomes. Second, we describe possible selection models which characterize the baseline and time-dependent factors which influence the timing of treatment. Together, the structural longitudinal model and the selection model determine the joint distribution of the observed longitudinal treatment and outcome data. Finally, given a causal target parameter and a framework for characterizing the full longitudinal data process we consider both parametric and semi-parametric methods of estimation and delineate the assumptions required for their validity.

2 Models for Longitudinal Outcomes and Exposures

To consider estimation approaches for longitudinal data subject to non-compliance, we first outline a class of causal models for the longitudinal outcomes, and then present a hierarchy of assumptions for factors that influence changes in exposure over time. By combining structural mean models with standard random effects assumptions the longitudinal outcome model allows identification of the causal treatment comparisons of interest and provides a complete likelihood for the observed data.

2.1 Outcomes: Structural Models for Longitudinal Data

Often the primary goal of medical research is to determine whether one or more treatments cause improvement in a medical condition. One way to define a causal effect of treatment is via potential outcomes (Greenland *et al.*, 1999). That is, if the outcome of interest is Y, then one can define a potential outcome $Y_i(x_k)$ for each subject i under each possible treatment x_k. Then, for subject i, a causal effect comparing treatment x_k to the referent treatment x_0 is $Y_i(x_k) - Y_i(x_0)$. Unfortunately this causal effect cannot in general be directly observed because each subject receives only one treatment. However, it is well known that the ACE can be estimated by randomizing subjects to the treatment(s) of interest or to a referent group, and then comparing the average outcome in each treatment group to the average outcome in the referent group. That is, one can calculate $E_i[Y(x_k)] - E_j[Y(x_0)]$ where subjects i=1,...,I receive treatment x_k and subjects j=1,...,J receive the referent treatment x_0.

With longitudinal data, definition of causal effects should consider the repeated exposure and outcome measurements that are potentially observed. If response measurements are made only at one pre-defined timepoint, then the outcome of interest will be defined as before, but treatment may become a vector of treatments over defined intervals prior to the outcome measurement. That is, one must consider the entire treatment path $\mathbf{x_k}$ that represents the vector of treatment indicators at each time t where $t = 1, 2, ..., T$: $\mathbf{x_k} = (x_{k1}, x_{k2}, ..., x_{kT})$. Then the ACE can be estimated via $E_i[Y(\mathbf{x_k})] - E_j[Y(\mathbf{x_0})]$ where subjects i=1,...,I follow treatment path $\mathbf{x_k}$ and subjects j=1,...,J follow the referent treatment path $\mathbf{x_0}$.

When outcome data is also collected at several different points in time, each subject has a vector of potential outcomes, one at each observation time t, under every possible treatment path: $\mathbf{Y_i}(\mathbf{x_k}) = [Y_{i1}(\mathbf{x_k}), Y_{i2}(\mathbf{x_k}), ..., Y_{iT}(\mathbf{x_k})]$. In this manuscript we restrict discussion to "point treatments" that occur at a single time s and where we assume subjects remain in the treated group at all times after s since this reflects the structure of a surgical intervention. Note that s can be any time during the observation period. In this situation the vector of potential outcomes can be denoted $\mathbf{Y_i}(s)$ and the individual causal effect at time t is $Y_{it}(s = 0) - Y_{it}(s = \infty)$ which contrasts the outcome at time t when a subject has surgery immediately after baseline ($s = 0$) to the outcome when surgery is withheld ($s = \infty$, or alternatively $s > T$). The corresponding ACE is $E[Y_{it}(s = 0) - Y_{it}(s = \infty)] = \Delta_s$.

Characterization of individual and average causal effects requires an individual-level, time-specific causal model that compares a person's potential outcomes under the two treatment paths of interest at any given time. We propose a longitudinal structural mixed model (LSMM) that contains three components: a group average, a subject average, and individual observations. This model is "structural" because its coefficients represent causal effects based on potential outcomes (Greenland *et al.*, 1999), and it is "mixed" because it incorporates both fixed and random effects to account for the correlation between repeated (potential) measurements on the same subject. In the context of a monotone treatment such as surgery, the LSMM can be written as follows:

$$\begin{aligned} Y_{it}(s) &= \mathbf{X}_i(t,s)'\boldsymbol{\beta} && \text{group average} \\ &+ \mathbf{Z}_i(t,s)'\mathbf{b}_i && \text{subjects within a group} \\ &+ e_{it}(s) && \text{observations within a subject} \end{aligned}$$

In this context, t is the time of measurement and s is the time of surgery. The group average component is typically the one of interest, and is often defined to be

$$\mathbf{X}_i(t,s)'\boldsymbol{\beta} = \mathbf{L}_i(t)'\boldsymbol{\lambda} + \mathbb{1}_{[t \geq s]} \cdot \mathbf{G}_i(t,s)'\boldsymbol{\gamma}$$

with $\mathbf{L}_i(t)'\boldsymbol{\lambda}$ representing the average outcome over time when no treatment occurs and $\mathbf{G}_i(t,s)'\boldsymbol{\gamma}$ representing the change in outcome attributable

to surgery, which may depend on the time at which surgery occurs. The subject-specific latent effects, represented by $\mathbf{b}_i$, account for the correlation between measurements on the same individual by assuming that each subject follows their own trajectory. Random effects are assumed to be independent of the measurement errors $e_{it}(s)$. A rank-preserving model is one in which the rank of individuals' outcomes are preserved across all possible treatment paths (Hernan *et al.*, 2005). For example, if subject 1 will have a higher outcome than subject 2 in the absence of treatment, then by a rank-preserving assumption he will also have a higher outcome in the presence of treatment. If $e_{it}(s)$ does not depend on the surgery time and treatment effects are assumed to be homogeneous, then the LSMM is rank-preserving. A non-rank-preserving model, which is often considered to be more realistic, can be constructed either by allowing $e_{it}(s)$ to differ before and after surgery occurs or by allowing treatment effects to be heterogeneous via a random effect.

2.2 Exposure: Statistical Models for Endogenous Treatment Process

For longitudinal data with noncompliance, the structural model alone is not sufficient for estimation of average causal effects. The exposure process, i.e. characterization of who receives treatment and when they receive it, must also be considered. If treatment is determined solely by factors that are separate from the longitudinal outcomes, then the treatment process is said to be exogenous. Specifically, given all previous exposures and time-invariant covariates, the exposure at time t is exogenous if it is independent of all preceding outcome measurements (Diggle *et al.*, 2002). One example of such an exogenous exposure is treatment in a randomized trial with perfect compliance. For example, in a surgery trial, surgery is exogenous if the subjects assigned to control never undergo surgery and the subjects assigned to surgery undergo surgery immediately after enrollment. In this case, the timing of surgery would only occur at enrollment so that outcomes after enrollment would have no bearing on treatment received.

However, in the case of an unblinded trial with crossover between treatment arms, the treatment received is often influenced by patient factors prior to randomization or, in the case of longitudinal trials, by interim patient factors including interim outcomes. That is, if a patient is assigned to the control group, but has poor outcomes as measured prior to treatment, then he may undergo surgery after these measurements despite having been assigned to non-surgical therapy. Any such treatment process that does not meet the condition for exogeneity is said to be endogenous.

One way that such endogeneity can arise is if previous outcomes directly determine whether or not a subject receives treatment at a subsequent time. An example of such "direct" selection is when a subject undergoes surgery as a result of having had a poor outcome at earlier times and then

actively seeking out a change in treatment. However, endogeneity could also be the result of an "indirect" selection process where unmeasured factors drive treatment choice. One possible indirect selection mechanism is when an underlying, unmeasured patient characteristic impacts both treatment status and outcome. For example, rather than a subject undergoing surgery because his last outcome was poor, he may undergo surgery because his physician thinks that he is likely to benefit from it, based on unmeasured aspects of available clinical information such as imaging data. Choosing patients selectively based on their propensity to benefit, which can be thought of as an unmeasured confounder, is a form of selection bias since not all of the target subjects are treated. If we could measure this latent propensity for benefit and condition on it, then treatment and outcome would be conditionally independent of each other.

We allow the selection model for p_{it}, which is the probability of subject i being treated at time t to include both direct and indirect selection by using the following model: $\text{logit}(p_{it}) = \alpha(t, R_i) + \eta(\bar{Y}_{i\,t-1}) + \delta(\mathbf{b}_i)$, where R_i is 0 if subject i is assigned to the control group and 1 if subject i is assigned to the surgery group. Simpler selection models can be constructed by setting one or more of the terms in this model equal to zero.

The observed data likelihood for longitudinal data with an endogenous exposure process must incorporate both this selection model and the LSMM specified in the previous section. In order to validly estimate the average causal effect when the exposure is endogenous, assumptions must be made about the underlying selection model. Analysis options and their corresponding assumptions will be described in the next section.

3 Estimation: Inference with Time-dependent Covariates

Standard estimation methods used to analyze longitudinal data will not necessarily be valid in the presence of endogenous treatment exposure (Diggle *et al.*, 2002). In this section we discuss both the assumptions needed for validity of standard estimators, as well as alternate estimators that relax key assumptions. Alternate estimators include maximization of more complex joint likelihoods, as well as methods tailored to estimate causal effects in the presence of endogeneity, such as marginal structural models, instrumental variable estimators, and g-estimators. In the subsections that follow we overview the general approach of each candidate method and emphasize key assumptions that are required for valid application.

3.1 Joint Likelihood of Outcomes and Treatment

LME and GEE estimators are not necessarily valid when exposure is endogenous (Diggle *et al.*, 2002). One way to incorporate the endogeneity

of exposure is to maximize the joint likelihood of observed outcomes and treatment under explicit assumptions about 1) selection for treatment and 2) the underlying structure of treatment effects.,

General Approach: To express the joint model mathematically, we need notation for the observed outcomes, as opposed to the counterfactual outcomes that we have considered thus far. Let each observed outcome be Y_{it}^O, let random effects remain denoted as $\mathbf{b}_i$, and now let W_{it}^O be an indicator of whether subject i has received treatment by time t, i.e. whether $t \geq s$. The vectors of observations for each subject will be denoted $\mathbf{Y}_i^O$ and $\mathbf{W}_i^O$, and the subsets of these vectors that represent histories through time t will be denoted $\bar{\mathbf{Y}}_{it}^O$ and $\bar{\mathbf{W}}_{it}^O$. In general we want to maximize the joint likelihood $\prod_i [\mathbf{Y}_i^O, \mathbf{W}_i^O]$. Using a telescoping factorization over time, the likelihood can be rewritten as:

$$\begin{aligned}
&\prod_i \int [\mathbf{Y}_i^O, \mathbf{W}_i^O | \mathbf{b}_i] dF(\mathbf{b}_i) \\
&= \prod_i \int \prod_t [Y_{it}^O | \bar{\mathbf{Y}}_{i\,t-1}^O, \bar{\mathbf{W}}_{it}^O, \mathbf{b}_i][W_{it}^O | \bar{\mathbf{Y}}_{i\,t-1}^O, \bar{\mathbf{W}}_{i\,t-1}^O, \mathbf{b}_i] dF(\mathbf{b}_i). \quad (1)
\end{aligned}$$

If we assume conditional independence of the outcomes, given treatment and random effects, then the joint likelihood becomes:

$$\prod_i \int \prod_t [Y_{it}^O | \bar{\mathbf{W}}_{it}^O, \mathbf{b}_i][W_{it}^O | \bar{\mathbf{Y}}_{i\,t-1}^O, \bar{\mathbf{W}}_{i\,t-1}^O, \mathbf{b}_i] dF(\mathbf{b}_i). \quad (2)$$

This likelihood is the most general formulation of the joint model that we will consider in this manuscript. The underlying causal assumptions appear in the first part of the likelihood $[Y_{it}^O | \bar{\mathbf{W}}_{it}^O, \mathbf{b}_i]$, while the selection model appears in the second part $[W_{it}^O | \bar{\mathbf{Y}}_{i\,t-1}^O, \bar{\mathbf{W}}_{i\,t-1}^O, \mathbf{b}_i]$. Given assumptions about each of these models, the likelihood can be maximized to estimate the target parameters in the first part, which will represent average causal treatment effects.

The general joint likelihood differs from the likelihood maximized by standard LME methods because it allows the selection model to include shared random effects. If additionally we assume that treatment depends only on previous outcomes and not on random effects, then we can simplify the selection model and obtain:

$$\prod_i \int \prod_t [Y_{it}^O | \bar{\mathbf{W}}_{it}^O, \mathbf{b}_i][W_{it}^O | \bar{\mathbf{Y}}_{i\,t-1}^O, \bar{\mathbf{W}}_{i\,t-1}^O] dF(\mathbf{b}_i)., \quad (3)$$

Now the simplified joint likelihood can be factored to allow separate consideration of the structural model and the selection model:

$$\prod_i \left\{ \left[\prod_t [W_{it}^O | \bar{\mathbf{Y}}_{i\,t-1}^O, \bar{\mathbf{W}}_{i\,t-1}^O] \right] \left[\int \prod_t [Y_{it}^O | \bar{\mathbf{W}}_{it}^O, \mathbf{b}_i] dF(\mathbf{b}_i) \right] \right\}. \quad (4)$$

Factorization permits estimation of structural parameters by separately maximizing $\prod_i \left[\prod_t \int [Y_{it}^O | \bar{\mathbf{W}}_{it}^O, \mathbf{b}_i] dF(\mathbf{b}_i)\right]$. LME models are fitted via maximization of $\prod_i \int [\mathbf{Y}_i^O | \mathbf{W}_i^O, \mathbf{b}_i] dF(\mathbf{b}_i)$. Therefore LME models will give consistent estimates of treatment effect when the joint likelihood can be rewritten as in (4) because this likelihood takes the same form as the LME likelihood. ,

Key Assumptions: In the presence of an endogenous treatment process, LME estimators, which do not explicitly incorporate a selection model, can be biased unless selection is based only on previous outcomes and not on shared latent effects. If selection does not depend on an indirect latent effect, then the LME estimator will provide a consistent estimate of treatment effects provided that the random effect structure is correctly specified and that there is no serial correlation among outcomes. Under these same conditions, GEE estimates will only be consistent in the special case where random intercepts accurately captures the within-subject dependence structure and GEE is implemented via an exchangable working covariance matrix. In this case GEE estimation is asymptotically equivalent to an LME estimator that assumes random intercepts. When there is an indirect latent effect and/or there is serial correlation between outcomes, LME and GEE estimates will both be biased. Such indirect selection is comparable to the nonignorable random-coefficient-based dropout described by Little in the context of missing data (Little, 1995).

When indirect selection exists, the joint likelihood that explicitly incorporates such selection should be the focus of likelihood-based estimation. Direct maximization of this joint likelihood will give consistent estimates, provided dependence on $\mathbf{b}_i$ for selection is correctly specified. An analytic solution to this maximization problem may be possible, but would require specialized numerical methods. A simple alternative to pure likelihood analysis would be to adopt Bayesian analysis that explicitly incorporates the selection model. For example, the posterior distribution corresponding to the likelihood in equation (2) is proportional to

$$\prod_i \prod_t [Y_{it}^O | \bar{\mathbf{W}}_{it}^O, \mathbf{b}_i; \boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2][W_{it}^O | \bar{\mathbf{Y}}_{i\,t-1}^O, \bar{\mathbf{W}}_{i\,t-1}^O, \mathbf{b}_i; \boldsymbol{\alpha}][\mathbf{b}_i; D] \cdot \pi(\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2, \boldsymbol{\alpha}, D).$$

Assuming independence between components of the prior distribution, it can be decomposed into $\pi(\boldsymbol{\beta}, \boldsymbol{\delta}) \cdot \pi(\sigma^2) \cdot \pi(\boldsymbol{\alpha}) \cdot \pi(D)$. Bayesian estimation can then be implemented in freely-available software, such as WinBUGS (Lunn *et al.*, 2000).

3.2 Marginal Structural Models

The likelihood-based methods discussed in Section 3.1 rely on the assumption that the selection model depends only on variables that are included in the structural model such as past outcomes and shared latent effects.

However, there are scenarios where additional time-dependent patient characteristics contribute to treatment decisions, yet the additional variables may be intermediate outcomes and therefore should not be included as covariates in the structural model. For example, in the SPORT example, when the outcome of interest is patient function as measured by the SF-36 physical function subscale, a subjects' pain score may contribute to treatment decisions and may be correlated with function measures. Including lagged pain scores as covariates in the structural model would likely attenuate the estimated effect of treatment on physical function since any treatment effects mediated by previous changes in pain would be controlled for in such a regression model. Therefore, methods that can more flexibly account for the various patient and/or provider factors that can influence treatment timing is an important generalization for non-compliance analysis.,
General Approach: Marginal structural models (MSMs) require the specification of a separate selection model that can include any covariates that are predictive of treatment. For example, with a time-varying dichotomous treatment, the selection model could be (Robins *et al.*, 2000)

$$\text{logit P}[W_t^O = 1|\bar{\mathbf{W}}_{t-1}^O = \bar{\mathbf{w}}_{t-1}, \bar{\mathbf{L}}_t = \bar{\mathbf{l}}_t] = \alpha_0 + \alpha_1 \cdot t + \alpha_2 \cdot w_{t-1} + \alpha_3 \cdot l_t + \alpha_4 \cdot w_{t-1} \cdot l_t$$

where $\mathbf{L}$ is a vector of all covariates that predict longitudinal treatment variables, which may or may not include previous outcome measurements $\bar{\mathbf{Y}}_{t-1}^O$. The selection model could also be a more complicated function of covariate history, but whatever model is deemed to be appropriate can be used to calculate the probability that each subject received his/her own treatment, conditional on past treatment and observed covariates, i.e.

$$p_{it} = \text{P}[W_t^O = 1|\bar{\mathbf{W}}_{t-1}^O = \mathbf{0}, \bar{\mathbf{L}}_t = \bar{\mathbf{l}}_t] \text{ or } (1-p_{it}) = \text{P}[W_t^O = 0|\bar{\mathbf{W}}_{t-1}^O = \mathbf{0}, \bar{\mathbf{L}}_t = \bar{\mathbf{l}}_t] \ .$$

Weighting the data by the inverse of these estimated probabilities creates a pseudopopulation in which the average causal effect is no longer confounded by selection bias. However, the weights can be highly variable, making the estimate of the causal effect highly imprecise, so a stabilized version of the weights is recommended (Robins *et al.*, 2000). In addition to the probabilities p_{it}, stabilized weights (sw_{it}) also require a model for the probability that each subject received his/her own treatment, conditional only on past treatment and baseline covariates, i.e.

$$p_{it}^* = \text{P}[W_t^O = 1|\bar{\mathbf{W}}_{t-1}^O = \mathbf{0}] \text{ or } (1-p_{it}^*) = \text{P}[W_t^O = 0|\bar{\mathbf{W}}_{t-1}^O = \mathbf{0}] \ .$$

The probabilities p_{it} and p_{it}^* are used to calculate the weights:

$$sw_{it} = \frac{\prod_{k=0}^{t} (p_{ik}^*)^{w_{ik}} (1-p_{ik}^*)^{(1-w_{ik})}}{\prod_{k=0}^{t} (p_{ik})^{w_{ik}} (1-p_{ik})^{(1-w_{ik})}}.$$

Once the weights have been specified, then a weighted GEE analysis of the structural model discussed in Section 2.1, using independent working covariance, will provide consistent estimates of the causal parameters of interest (Robins *et al.*, 2000). ,

3.3 Instrumental Variables

An estimation approach popular in health services research exploits the properties of one or more instrumental variables (IVs) to eliminate the bias due to exposure selection, rather than explicitly modeling the selection process. An IV is a variable whose effects on the outcome are exerted only via the exposure of interest (Angrist *et al.*, 1996). Therefore any association between the IV and the outcome can be fully explained by exposure. One common example of an IV is randomization status - it determines exposure, but in most scenarios will have no other impact on outcomes. ,
General Approach: Causal inference using IVs can be understood in the context of a system of equations. The first equation is the structural model of interest, and the second equation uses one or more IVs to predict exposure (Hogan and Lancaster, 2004), e.g.

$$Y_i = \lambda + \gamma W_i + \epsilon_i \quad (5)$$

$$W_i = \zeta_0 + \zeta_1 R_i + \delta_i \quad (6)$$

where δ_i has mean zero and can depend on ϵ_i. The predicted exposure from equation (6) represents the expected exposure that is due to the IV(s). By substituting this expected exposure into the structural equation (5), one can recover the effect of exposure on outcome in the absence of selection because only the exposure that is due to the IV(s) is included. This approach is called two-stage least squares (2SLS).
In the context of longitudinal data, one or both sets of model parameters can be estimated via GEE or LME estimators to account for the correlation between outcomes collected at different timepoints (Bond *et al.*, 2007). Alternatively, 2SLS can be cast as a method of moments (MM) or generalized method of moments (GMM) estimator, depending on the number of IVs (Wooldridge, 2002). Econometricians have developed tools to implement these GMM methods in the presence of random intercepts (Stata Press, 2007). These implementations are comparable to 2SLS with a LME outcome model, but they use a different weighting scheme. ,

3.4 G-estimation

A semiparametric method for estimating average causal effects in the presence of time-dependent confounding uses g-estimation for the parameters of a structural nested model (SNM) (Robins, 1994). G-estimation relies on estimation of treatment-free potential outcomes for all subjects who received treatment. These treatment-free outcomes represent the outcomes that each subject would have had if they had received no treatment instead of the treatment that they actually received. ,
General Approach: As in Robins' work, let the vector of treatment-free outcomes for subject i at times t=1,...,K be written as $\bar{\mathbf{H}}_\mathbf{i}(\boldsymbol{\psi_0})$ where $\boldsymbol{\psi_0}$ represents the true values of the average causal effects $\psi_{0,1}$ and $\psi_{0,6}$ (Robins,

$\bar{\mathbf{q}}_{opt,i}$	$E_{R_i}[H_i(\psi)\|Y_{i0}]$
$\mathbf{g}_{opt,i}$	$\mathbf{w_i}\{\mathbf{D_i} - E_{R_i,Y_{i0}}[\mathbf{w_i}]^{-1}E_{R_i}[\mathbf{w_i}\mathbf{D_i}\|Y_{i0}^O]\}$
$\mathbf{D_i}$	$E_{\mathbf{a}}[\frac{\partial \bar{\mathbf{H}}_\mathbf{i}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}'}\|R_i, Y_{i0}]$
$\mathbf{w_i}$	$\{E_{\mathbf{a}}[(\bar{\mathbf{H}}_\mathbf{i}(\boldsymbol{\psi}) - \bar{\mathbf{q}}_{opt})(\bar{\mathbf{H}}_\mathbf{i}(\boldsymbol{\psi}) - \bar{\mathbf{q}}_{opt})'\|R_i, Y_{i0}]\}^{-1}$

TABLE 1. Functions needed to implement semi-parametric efficient g-estimation.

1994). For subjects who were never treated, $\bar{\mathbf{H}}_\mathbf{i}(\boldsymbol{\psi_0})$ is equivalent to the vector of observed outcomes. For those who were treated, it is the vector of observed outcomes until the treatment occurs, after which it is the vector of observed outcomes minus the causal effect of treatment.
Due to randomization and the assumption of no-current-treatment-interaction , the average treatment-free outcomes in the group assigned to control should be the same as those in the group assigned to treatment, leading to the following estimating equations: $n^{-1/2}\sum_i \mathbf{d}'(R_i, Y_{i0}^O)[\bar{\mathbf{H}}_\mathbf{i}(\boldsymbol{\psi}) - \bar{\mathbf{q}}(Y_{i0}^O)]$ where $\mathbf{d}$ and $\bar{\mathbf{q}}$ are arbitrary functions except that $\mathbf{d}$ must have the form $\mathbf{g}(R_i, Y_{i0}^O) - E[\mathbf{g}(R_i, Y_{i0}^O)|Y_{i0}^O]$. Note that Y_{i0}^O includes all observed baseline covariates for subject i, but we have used this notation because outcome will be the only baseline variable included in our simulations. Due to Proposition 1 in section 3 of Chamberlain's paper (Chamberlain, 1992), these estimating equations will be semiparametric efficient with the "optimal" choices of functions given in Table 1. The choice of $\mathbf{g}_{opt}$ results in $E[\mathbf{g}(R_i, Y_{i0}^O)|Y_{i0}^O] = 0$, so that $\mathbf{g}_{opt}$ essentially replaces $\mathbf{d}$ in the estimating equations. Note that we have replaced Robins' $\boldsymbol{\mu}$ with $\mathbf{D}$ in order to clarify that this quantity functions like a derivative in general estimating equation theory. Regardless of the choices of $\bar{\mathbf{q}}_i$ and $\mathbf{g}_i$, the estimate of $\boldsymbol{\psi}$ and its variance would then be (Goetghebeur and Lapp, 1997):

$$\begin{aligned} \hat{\boldsymbol{\psi}} &= (\mathbf{d}'\mathbf{M})^{-1}\mathbf{d}'(\mathbf{Y} - \bar{\mathbf{q}}) \\ \mathrm{Var}[\hat{\boldsymbol{\psi}}] &= (\mathbf{d}'\mathbf{M})^{-1}\frac{1}{n}\sum_i \{\mathbf{d}_i'(\bar{\mathbf{H}}_i - \bar{\mathbf{q}}_i)(\mathbf{d}_i'(\bar{\mathbf{H}}_i - \bar{\mathbf{q}}_i))'\}(\mathbf{d}'\mathbf{M})'^{-1} \end{aligned}$$

where $\mathbf{Y}$ and $\mathbf{q}$ are column vectors consisting of stacked vectors of post-baseline observations and $\bar{q}_i$ for each subject, $\mathbf{d}$ is the matrix consisting of stacked subject-specific matrices $\mathbf{d}_i$ each with the number of rows equal to the number of post-baseline observations and the number of columns equal to the number of treatment effects to be estimated, and $\mathbf{M}$ is the design matrix associated with the causal effects, also consisting of stacked subject-specific design matrices each with the same dimensions as $\mathbf{d}_i$.
The use of $\mathbf{D}$ in the simple case where we assume constant $\mathbf{w}$ is equivalent to the use of expected exposure from the first-stage model of 2SLS IV

	Estimator					
	GEE	LME	JOINT	MSM	IV	G-EST
Implemented in standard software	yes	yes	no	yes	yes	no
Requires correctly specified structural model	yes	yes	yes	yes	yes	yes
Requires correctly specified selection model	no	no	yes	yes	no	no
Consistent with direct selection	no	yes	yes	yes	yes	yes
Consistent with indirect selection	no	no	yes	no	yes	no
Permits selection model and structural model to include different covariates	no	no	yes	yes	no	no

TABLE 2. Properties of methods for estimating causal effects in longitudinal data.

procedures. That the g-estimator reduces to an IV estimator under these assumptions has been observed previously (Joffe and Brensinger, 2003; Dunn and Bentall, 2007). Thus although g-estimation would likely permit more efficient estimation of the average causal effect, its application in our scenario of interest is left for future work, and only a simpler version of it - the IV estimator - will be illustrated here. ,

4 Summary of Estimation Options

The properties of the methods presented in this section are summarized in Table 2. All methods except joint likelihood maximization and g-estimation have been implemented in standard software. Joint likelihood maximization can be implemented in WinBUGS. To our knowledge, a general implementation of semiparametric-efficient g-estimation has not been distributed, except with simplifying assumptions that are not satisfied in our problem of interest.

Estimation of the average causal effect of treatment is a primary goal of randomized clinical trials. We have considered a scenario in which longitudinal data is available and interest lies in a one-time treatment, such as surgery, that is administered with noncompliance to randomized treatment assignment. The causal effect to be estimated can be specified by constructing a LSMM for the potential outcomes. The formulation of a LSMM is useful because it clearly identifies the effect of interest and is compatible with a number of different estimation methods. In choosing an estimation method, both standard longitudinal methods and methods designed specifically to estimate causal effects should be considered.

References

Angrist JD, Imbens GW and Rubin DB. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, **91** (434), 444–455.

Bond, SJ, White, IR, and Walker, AS. (2007). Instrumental Variables and interactions in the causal analysis of a complex clinical trial. *Statist Med*, **26** (7),1473–1496.

Chamberlain, G. (1992). Efficiency Bounds for Semiparametric Regression. *Econometrica*, **60** (3), 567–596.

Diggle, PJ, Heagerty, P, Liang, KY, and Zeger, SL. (2002). *Analysis of Longitudinal Data*, second edition. Oxford University Press, Oxford, UK.

Dunn, G and Bentall, R (2007). Modelling treatment-effect heterogeneity in randomized controlled trials of complex interventions. *Statist Med*, **26** (26),4719–4745.

Ellenberg, JH (1996). Intent-to-Treat Analysis Versus As-Treated Analysis. *Drug Inf J*, **30**, 535–544.

Flum, DR (2006). Interpreting Surgical Trials With Subjective Outcomes: Avoiding UnSPORTsmanlike Conduct. *Journal of the American Medical Association*, **296** (20), 2483–2485.

Goetghebeur, E and Lapp, K (1999). The Effect of Treatment Compliance in a Placebo-controlled Trial: Regression with Unpaired Data. *Appl Statist*, **46** (3),351–364.

Greenland, S, Robins, JM, and Pearl, J (1999). Confounding and Collapsibility in Causal Inference. *Stat Sci*, **14** (1), 29–46.

Hernan, MA, Cole, SR, Margolick, J, Cohen, M, and Robins, JM (2005). Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidem Dr S*, **14** (7), 477–491.

Hogan, JW and Lancaster, T (2004). Instrumental Variables and Inverse Probability Weighting for Causal Inference from Longitudinal Observational Studies. *Stat Meth Med Res*, **13** (1), 17–48.

Joffe, MM, and Brensinger, C (2003). Weighting in instrumental variables and G-estimation. *Statist Med*, **22** (8), 1285–1303.

Little, RJA (1995). Modeling the Drop-Out Mechanism in Repeated-Measures Studies. *Journal of the American Statistical Association*, **90** (431), 1112–1121.

Lunn, DJ, Thomas, A, Best, N, and Spiegelhalter, D (2000). WinBUGS a Bayesian modelling framework: concepts, structure, and extensibility. *Stat Comput*, **10** (4), 325–337.

Robins, JM (1994). Correcting for Noncompliance in Randomized Trials using Structural Nested Mean Models. *Commun Statist, Theory Meth*, **23** (8), 2379–2412.

Robins, JM, Hernan, MA, and Brumback, B (2000). Marginal Structural Models and Causal Inference in Epidemiology. *Epi*, **11** (5), 550–560.

Stata Press (2007). *Longitudinal/Panel Data Reference Manual*, Release 10. StataCorp LP, College Station, TX.

Weinstein, JN, Tosteson, TD, Lurie, JD, Tosteson, ANA, Hanscom, B, Skinner, JS, Abdu, WA, Hilibrand, AS, Boden, SD, and Deyo, RA (2006). Surgical vs Nonoperative Treatment for Lumbar Disk Herniation: The spine patient outcomes research trial (sport) randomized trial. *Journal of the American Statistical Association*, **296** (20), 2441–2450.

Wooldridge, JM (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA.

Part 2. Contributed papers

Fitting the therapy term in a Gompertz diffusion process

G. Albano[1], V. Giorno[2], P. Román-Román[3], F. Torres-Ruiz[3]

[1] Dipartimento di Scienze Economiche e Statistiche. Università di Salerno, 84084 Fisciano (SA), Italia. *pialbano@unisa.it*
[2] Dipartimento di Matematica e Informatica. Università di Salerno. 84084, Fisciano (SA), Italia. *giorno@unisa.it*
[3] Departamento de Estadística e Investigación Operativa. Universidad de Granada. 18071, Granada, España. {*proman,fdeasis*}*@ugr.es*

Abstract: We consider a Gompertz-type diffusion process including, in its infinitesimal mean, a time-dependent function representing a *therapy* that can modify the behavior of the variable under study. A strategy is proposed in order to *fit* this function, if it is unknown. An application to tumor growth is presented.

Keywords: Gompertz diffusion process; Tumor growth; Therapy.

1 Introduction

The Gompertz curve is associated with the study of populations in which one can observe an initially slow growth, followed by a period of rapid growth in the variable under consideration and by another period where the growth speed decreases as the variable approaches a limit value. This value, the carrying capacity, represents the limitation of the natural resources. As regards the biomedical applications, in the sixties Laird (1964) used successfully for the first time this curve to fit data of growth of tumors. In fact, tumors are cellular populations growing in a confined space where the availability of nutrients is limited.
In the last decades many efforts have been oriented in order to build stochastic models associated with this curve. The main aim is to take into account fluctuations or disturbances, not always quantifiable or even unknown, that might exist in the system under consideration and not considered by the usual deterministic models. Stochastic models have been built by means of stochastic differential equations whose solutions are diffusion processes. In this sense, Gutiérrez et al. (2007) introduced a diffusion process with the aim of obtaining a continuous stochastic model associated with the Gompertz curve whose limit value depends on the initial one. This process is a particular case of the lognormal diffusion process with exogenous factors, whose main characteristics are summarized in Gutiérrez

et al. (2006). The introduction of exogenous factors can be very interesting as regards to the application of this process to real data. For example, these exogenous variables can model the effect of an epidemic in a population or, in a study over tumor growth, they can represent the introduction of a therapy affecting the growth of tumoral cells. This last question was studied by Albano and Giorno (2006), but the process considered in that case is that associated with an expression of the Gompertz curve whose limit value does not depend on the initial one.
In this paper we consider a modification of the former Gompertz diffusion process abovementioned by adding a continuous function in the infinitesimal mean that represents a *therapy* included in the system. The knowledge of the functional form of this function is fundamental because of allows to introduce an external control to the system under consideration. On other hand, the study of some questions related to the process, for example the first-passage-time problem through a boundary, needs the functional form of this exogenous variable, question that is not always possible. For this reason, and for such a case, in the following section we suggest a strategy for *fitting* the functional form of the therapy. Later we show the possibilities of this method with an application to tumor growth data.

2 The model: fitting the therapy

The Gompertz process introduced by Gutiérrez et al. (2007) is a diffusion process $\{X(t) : 0 \leq t_0 \leq t \leq T\}$ taking values on R^+ and with infinitesimal moments given by $A_1(x,t) = m\mathrm{e}^{-\beta t}x$ and $A_2(x,t) = \sigma^2 x^2$, where $m \in R$, $\beta > 0$ and $\sigma > 0$. This process is a particular case of the lognormal diffusion process with exogenous factors, whose infinitesimal mean is of the form $h(t)x$ (h a continuous function in $[t_0, T]$). The main characteristics of this process are summarized in Gutiérrez et al. (2006). Concretely, the transition probability density is that of a lognormal variable $\Lambda_1(\gamma; \sigma^2(t-s))$ where $\gamma = y + \int_s^t h(\tau)\,d\tau - \frac{\sigma^2}{2}(t-s)$, $t > s$. On other hand, and taking an initial lognormal distribution ($X(t_0) \sim \Lambda_1(\mu_0; \sigma_0^2)$), the finite dimensional distributions are lognormal (we notice that it is possible to choose a degenerate distribution at t_0, being this choice a particular case of the former and ensuring also the lognormality of the finite dimensional distributions). In general, one concludes that the distribution of $X(t)$ is lognormal $\Lambda_1(\gamma; \sigma_0^2 + \sigma^2(t-t_0))$ where $\gamma = \mu_0 + \int_{t_0}^t h(s)\,ds - \frac{\sigma^2}{2}(t-t_0)$, $t > t_0$, from where $E[X(t)] = E[X(t_0)] \exp\left(\int_{t_0}^t h(s)\,ds\right)$. By taking $h(t) = m\mathrm{e}^{-\beta t}x$ in this expression, one can check that the mean function for the process $X(t)$ is a Gompertz curve with limit value depending on the initial one. On other hand, in order to introduce a *therapy* modifying the above trend, we consider a new lognormal diffusion process $\{X^C(t)\}$ by taking now as infinitesimal moments $A_1^C(x,t) = (m - C(t))\mathrm{e}^{-\beta t}x$ and $A_2^C(x,t) = A_2(x,t)$. Ob-

viously, the introduction of $C(t)$ (the therapy) causes an alteration in the infinitesimal growth rate of the variable modeled by the process, as well as in the trend of the process. Indeed, if one chooses the same initial distribution for both processes, one has $E\left[X^C(t)\right] = E\left[X(t)\right] \exp\left(-\int_{t_0}^{t} C(s)\mathrm{e}^{-\beta s}\, ds\right)$, showing, in term of $C(t)$, the difference between the process $X(t)$ (usually associated in practice with a control group) and $X^C(t)$ (related to a group where a therapy has been included). This expression motivates a procedure for fitting the therapy term if it was unknown. In such a case, and taking $m(t) = \log\left(E\left[X^C(t)\right]/E\left[X(t)\right]\right)$, it verifies $-m'(t)e^{\beta\, t} = C(t)$; so, in practice we can proceed as follows:

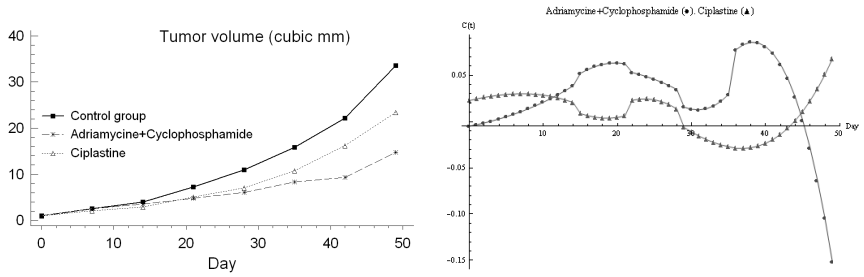
- For the control group, and from the observed values in times $t_1, \ldots, t_n$, to estimate the parameters of $X(t)$ as it is described in Gutiérrez et al. (2007). Let $\widehat{\beta}$ be the estimation of the parameter β.
- To approximate the values of $m(t)$ by $\log(x_i^c/x_i)$, where x_i is the mean value of the trajectories of the control group at t_i, whereas x_i^c is the corresponding in the group with therapy.
- To interpolate these values, to derive the interpolation function and to multiply the result by $e^{\widehat{\beta}\, t_i}$, $i = 1, \ldots, n$.

This procedure provides, for each time instant t_i, a value of the therapy, named c_i, from which we can approach the function $C(t)$ by means of some procedure (numerical interpolation, regression,...) in order to use it for estimating the mean of the process $X^C(t)$ (or others characteristics).

3 An application

In order to study the effect of two types of chemotherapies in breast cancer xenografts, we have considered data about the growth of $BC297MONp5$ from three experimental groups of mice. The first of them is a control group, whereas the others have received two treatments, Adriamycine-Cyclophosphamide and Ciplastine, respectively (see Figure 1a)). Firstly we have considered a process of the type $X(t)$ for the control group, being its parameters estimated using the procedure described in Gutiérrez et al. (2007). With these estimations we have considered a process of the type $X^C(t)$ for each one of the treatments, and following the strategy proposed in this paper, the functional form of the therapies have been fitted by numerical interpolation (see Figure 1b)). Finally, we have estimated the mean of the processes that include the therapies (see Figure 2).

Acknowledgments: The authors are very grateful to Dr. Didier Decaudin (Institut Curie, Paris) for providing the data here used. This work was supported in part by the Spanish grants MTM2008-05785, P06-FQM-02271



a) *b*)

FIGURE 1. a) Data for the three groups. b) Fitted therapies.

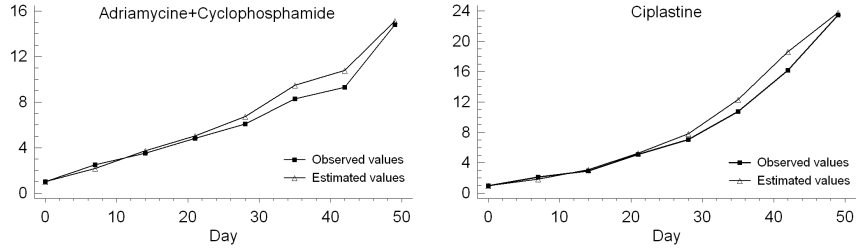


FIGURE 2. Estimated means under the therapies (mm^3).

and HI20007-0034, and by the italian grants PRIN 2008 and Azione Integrata Italia Spagna.

References

Albano, G., Giorno, V. (2006). A stochastic model in tumor growth. *Journal of Theoretical Biology*, **242**, 329–336.

Gutiérrez, R., Rico, N., Román, P., Torres, F. (2006). Approximate and generalized confidence bands for some parametric functions of the lognormal diffusion process with exogenous factors. *Scientiae Mathematicae Japonicae*, **64**, 843–859.

Gutiérrez, R., Román, P., Romero, D., Serrano, J.J., Torres, F. (2007). A new gompertz-type diffusion process with application to random growth. *Mathematical Biosciences*, **208**, 147–165.

Laird, A. K. (1964). Dynamics of tumor growth. *British Journal of Cancer*, **18**, 490-502.

Time Series and Mixed Models to Study the Country-Specific Outpatient Antibiotic Use in Europe

Girma Minalu Ayele[1], Niel Hens[1,2], Marc Aerts[1], Samuel Coenen[3], Arno Muller[3], Niels Adriaenssens[3], Philippe Beutels[2], Geert Molenberghs[1,4], Herman Goossens[3]

[1] Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BIOSTAT), Hasselt University, Belgium.
[2] Centre for Health Economics Research and Modeling Infectious Diseases Centre for the Evaluation of Vaccination (WHO Collaborating Centre) Vaccine & Infectious Disease Institute (VAXINFECTIO), Antwerp, Belgium.
[3] Laboratory of Medical Microbiology, Vaccine & Infectious Diseases Institute (VAXINFECTIO), University of Antwerp, Belgium.
[4] Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BIOSTAT), Catholic University of Leuven, Belgium.

Author for correspondence: Girma Minalu Ayele, I-BIOSTAT, Hasselt University, Campus Diepenbeek, Agoralaan 1, 3590, Diepenbeek, Belgium. E-mail: *girma.minaluayele@uhasselt.be*.

1 Introduction

Quarterly data on outpatient antibiotic use were collected from 27 European countries for the period 1997-2007 through the European project ESAC (www.esac.ua.ac.be). Antibiotic use was measured as defined daily doses per 1000 inhabitants per day (DID) for different Anatomical Therapeutic Chemical (ATC) levels, according to the WHO ATC classification. Community antibiotic use for 2005 in the four administrations of the UK was compared with other European countries. With a longitudinal analysis, data from 1997 to 2005 from the four administration of the UK and from Belgium was compared (Davey et al., 2008). In this abstract we focus on Tetracyclines. Since repeated measures were taken for the different countries, their correlation has to be taken into account when analyzing the data. Linear mixed models (Verbeke and Molenberghs, 2000) are used to assess country-specific trends over Europe, while accounting for country-specific global use as well as country-specific seasonal effects. These findings yield new important insights in the evolution of outpatient antibiotic use over several European countries.

2 Methods

Linear mixed models were applied (models 1-3) and extended by including sinusoidal components over time to account for the seasonal variation (models 4-5). Next to a country-specific random intercept, a country-specific random slope for time and for the sinusoidal component were included. The importance of the random variables (b_i) were assessed using likelihood ratio tests. AIC was used for model comparison.

$$DID_i = (\beta_0 + b_{0i}) + \beta_1 T + \varepsilon_i \quad (1)$$

$$DID_i = (\beta_0 + b_{0i}) + \beta_1 T + \beta_2 T^2 + \varepsilon_i \quad (2)$$

$$DID_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})\, T + \beta_2 Period_j + \varepsilon_{ij} \quad (3)$$

$$DID_i = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})\, T + \beta_2 \sin(\omega T + \delta) + \varepsilon_i \quad (4)$$

$$DID_i = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})\, T + (\beta_2 + b_{2i}) \sin(\omega T + \delta) + \varepsilon_i \quad (5)$$

For model (5), β_1 is the regression coefficient describing the marginal linear time trend (T) of the time series, β_2 is the amplitude, ω (in radians) is the frequency, δ (in radians) is the phase shift or phase angle, $b_i \sim N(0, D)$ where $b_i = (b_{0i}, b_{1i}, b_{2i})$ and $\varepsilon_i \sim N(0, \Sigma_i)$. Letting A=$\beta_2$sin($\delta$) and B=$\beta_2$cos($\delta$), it follows that B/A=tan(δ) and $\sqrt{A^2 + B^2}$ $= \beta_2 \sqrt{\cos^2(\delta) + \sin^2(\delta)} = \beta_2$, so phase angle and amplitude estimates can be constructed from estimates of A and B. Parameter estimates of A and B can be estimated directly from the linear mixed models (Brocklebank and Dickey, 2003).
A nonlinear mixed model was also fitted to the data, using this model all the variables in model (5) can be estimated directly. We used the MCMC procedure in SAS to fit MCMC-based mixed model. The following uninformative prior distributions were used for the random effects $b_i \sim N(0, D)$, a multivariate normal prior with mean vector 0 and variance matrix D (d_{ij}, i,j=1,2,3) with the simple covariance structure was used, gamma prior distributions were assigned to the $d_{ij}{}'s$.

3 Results and Conclusions

Country specific profiles of Tetracycline use (Figure 1) show considerable seasonal variation in all countries. For the final model, a model with the sinusoidal component (5), fixed effect parameters (standard error) were estimated to be 2.57(0.05), -0.01(2.E-3), 0.46(0.03) and 0.45(0.06) for intercept, linear time effect, amplitude and phase shift, respectively, using MCMC-based mixed model.

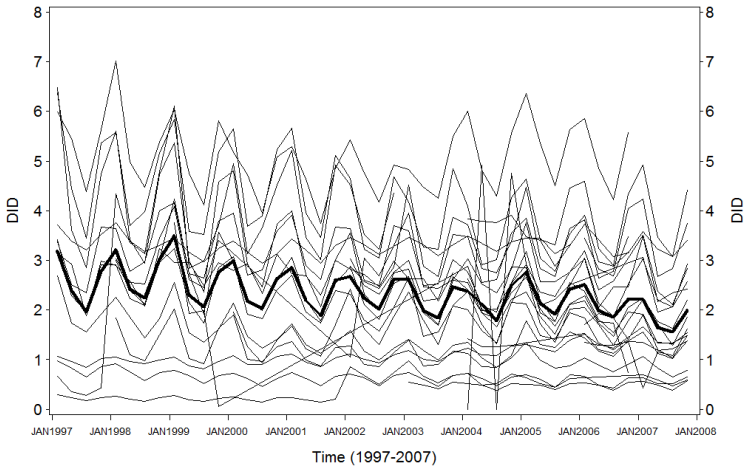


FIGURE 1. Country-specific evolutions and population averaged evolution (bold line) for Tetracycline use.

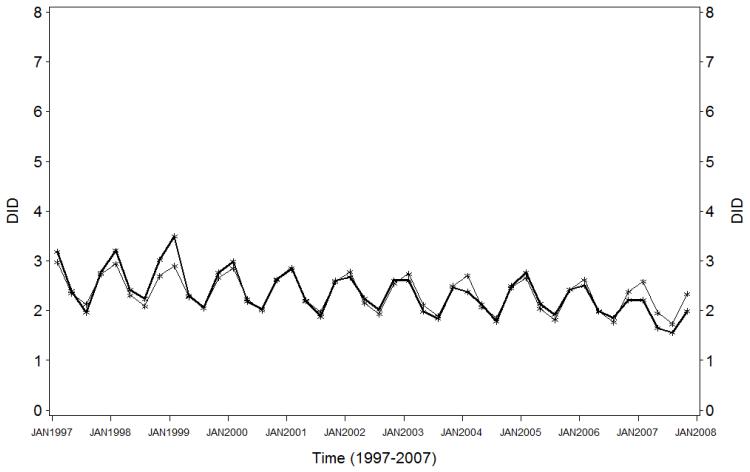


FIGURE 2. The predicted mean profile and population averaged evolution (bold line) for Tetracycline use.

Predicted mean profile and the overall mean profile (bold line) are shown in Figure 2. From the plot we can see the observed mean profile and the predicted mean profile using the MCMC-based mixed model are more close. Results of non-linear mixed model were compared to that obtained from a linear mixed model fitted to the same data. The nonlinear model fitted data better.

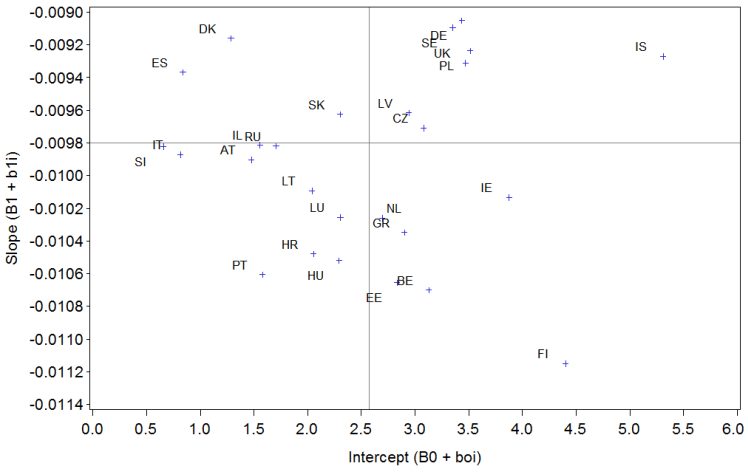


FIGURE 3. plot of slope for time (fixed effect + random effect) and plot of intercept (fixed effect + random effect) for Tetracycline use.

From the summery plot (Figure 3) we can see by how much the country-specific estimates are deviating from the overall estimates for linear time effect and intercept. The horizontal line (=-0.00980) is an estimate for the fixed linear time effect and the vertical line (=2.5746) is an estimate for the fixed effect intercept.
The correlation between the random effects was estimated to be -0.5732 (random intercept and random slope for time), 0.7683 (random intercept and random slope for sine) and -0.4465 (random slope for time and random slope for sine), respectively. Models were fitted and selected for each ATC, separately. For most ATCs there is a decrease in antibiotic use over time.
All country-specific trends for all ATCs (16 Anatomical Therapeutic Chemical levels) were examined through a multivariate cluster analysis, in order to identify a natural clustering of countries within Europe with similar trend patterns over all ATCs. Austria, Sweden, Germany, Lithuania, Russian Federation, Netherlands, United Kingdom, Latvia, Iceland and Croatia; Czech Republic, Finland and Spain; Israel and Portugal; Cyprus, Ireland and Estonia are clusters of countries with a similar trend.

References

Brocklebank, J.C., and Dickey, D.A.. (2003). *SAS for Forecasting Time Series, Second Edition*. Cary, NC: SAS Institute.

Davey, P., Ferech, M., Ansari, F., Muller, A. and Goossens, H. (2008). Outpatient antibiotic use in the four administrations of the UK: cross-sectional and longitudinal analysis. *JAC* **62**, 1441-1447.

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

Markov-switching autoregressive latent variable models for longitudinal data

Silvia Bacci[1], Francesco Bartolucci[2], Fulvia Pennoni[3]

[1] Department of Economics, Finance and Statistics, University of Perugia, Via A. Pascoli 20, 06123 Perugia, Italy, silvia.bacci@stat.unipg.it
[2] Department of Economics, Finance and Statistics, University of Perugia, Via A. Pascoli 20, 06123 Perugia, Italy, bart@stat.unipg.it
[3] Department of Statistics, University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy, fulvia.pennoni@unimib.it

Abstract: We propose a generalization of the autoregressive latent variable models for longitudinal data based on an AR(1) process to represent the effect of unobservable factors on the response variables. The generalization is based on assuming that the latent process follows a Markov-switching AR(1) process with correlation coefficient depending on the regime of the chain. Some particular cases are discussed in detail and illustrated by an application to a longitudinal dataset about self-evaluation of the health status.

Keywords: Latent Markov model; Ordinal variables; Numerical integration.

1 Introduction

In the analysis of longitudinal data, an important aspect to be taken into account is how to represent the effect that unobservable factors have on the occasion-specific response variables in addition to the effect of observable covariates. The simplest approach is based on the inclusion, in the model of interest, of individual-specific random intercepts. In this way, however, the effect of unobservable factors is assumed to be time constant. A natural way to relax this assumption is by assuming that, for each subject, there are occasion-specific random effects which follow an AR(1) process (Chi and Reinsel, 1989); the resulting model will be referred to as latent autoregressive (LAR) model. An alternative formulation is based on the inclusion of a sequence of discrete latent variables which follow a first-order Markov chain. In this way, a Latent Markov (LM) model (Wiggins, 1973) with covariate results. For a review on LM models see Bartolucci *et al.* (2010) and for an instance of a complex model formulated following this approach see Bartolucci and Farcomeni (2009).
The main advantage of the LAR formulation is that it retains a parsimony close to that of the corresponding random effect model. Moreover, in certain applications it is natural to represent the error terms by continuous

rather than discrete random variables. On the other hand, estimating the resulting model may be problematic from the computational point of view (Heiss, 2008). The model based on the LM formulation naturally provides a classification of subjects into a reduced number of groups, is easier to estimate, and may reach a better fit. However, this model is usually less parsimonious. It is also worth noting that a Markov chain is able to approximate adequately a continuous process and then the model based on the LM formulation may be seen as a semi-parametric version of the model based on the AR(1) process. The issue of the comparison between the two approaches above is related to that of the comparison between a standard random effect model and its latent class version in contexts simpler to the present one; see Lindsay *et al.* (1991) and Greene and Hensher (2003).
In this paper, we formulate a model for longitudinal data which is based on the assumption that the error terms follow a Markov-switching AR(1) process (Hamilton, 1989). In particular, we assume that a set of different regimes are possible, with each regime corresponding to a different value of the correlation coefficient. How a subject moves between regimes is governed by an unobservable Markov chain which is time-homogenous. In this way, we extend the LAR model by allowing the correlation coefficient to be different between subjects and occasions. Moreover, we expect that the resulting model has a fit comparable to that of a model based on a LM formulation, but it is more parsimonious. Two versions of the proposed model are discussed in detail. In the former, the autoregressive correlation coefficient may be different between subjects, but not between occasions. In the second version, instead, each subject randomly moves between different regimes. Both versions are estimated by the maximum likelihood method, which is implemented on the basis of an algorithm similar to the sequential numerical integration algorithm proposed by Heiss (2008).
The paper is organized as follows. In the following section we introduce the basic notation and describe the LAR formulation for longitudinal data. In Section 3 we outline the proposed extension, whereas the results of an illustrative application based on a dataset about self-evaluation of the health status are briefly illustrated in Section 4.

2 Preliminaries

Let y_{it} be the response variable observed at occasion $t = 1, \ldots, T$ for subject $i = 1, \ldots, n$ and let $\boldsymbol{x_{it}}$ be a corresponding vector of covariates.
The model based on the LAR formulation for these variables assumes that, for every subject i, $y_{i1}, \ldots, y_{iT}$ are conditionally independent given the covariates $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}$ and a sequence of latent variables $u_{i1}, \ldots, u_{iT}$ which follows an AR(1) process. In particular, we assume that $u_{i1} \sim N(0, \sigma^2)$ and

that, for $t > 1$,

$$u_{it}|u_{i,t-1} = \rho u_{i,t-1} + \varepsilon_{it}, \quad (1)$$

$$\varepsilon_{it} \sim N[0, \sigma^2/\sqrt{1-\rho^2}]. \quad (2)$$

An important point is how to model the conditional distribution of each response variable y_{it} given u_{it} and $\boldsymbol{x_{it}}$. For instance, in the case of ordinal responses with l categories, that we will consider in our application, a natural parametrization is based on cumulative (or global) logits:

$$\log \frac{p(y_{it} > j|u_{it}, \boldsymbol{x}_{it})}{p(y_{it} \leq j|u_{it}, \boldsymbol{x}_{it})} = \mu_j + u_{it} + \boldsymbol{x}_{it}'\boldsymbol{\beta}, \quad j = 1, \ldots, l-1.$$

The main difference with the LM formulation is that in the latter the latent process follows a Markov chain with k states, with the following parameters: $k-1$ support points (the first is fixed at 0), in addition to $k-1$ initial probabilities, and $k(k-1)$ transition probabilities. The LAR model uses, instead, only 2 parameters for the latent process.

3 The proposed model

We generalize the LAR model presented in the previous section in order to allow for a different correlation between time occasions and subjects. For this aim we exploit the general framework of Markov-switching autoregressive models (Hamilton, 1989), where the correlation coefficient depends on an unobserved Markov chain.

3.1 Model assumptions

The proposed model (named SW-LAR) is formulated as in Section 2, with assumptions (1) and (2) substituted by

$$u_{it}|u_{i,t-1}, v_{it} = \rho_{v_{it}} u_{i,t-1} + \varepsilon_{it},$$

$$\varepsilon_{it}|v_{it} \sim N[0, \sigma^2/\sqrt{1-\rho^2_{v_{it}}}],$$

where the latent process $v_{i1}, \ldots, v_{iT}$ follows a Markov-chain with k latent states corresponding to the correlation coefficients $\rho_1, \ldots, \rho_k$. This process is characterized by the vector of initial probabilities $\boldsymbol{\lambda}$, with elements λ_v, $v = 1, \ldots, k$, and the transition probability matrix $\boldsymbol{\Pi}$, with elements $\pi_{v_0 v}$, $v_0, v = 1, \ldots, k$. Note that every latent variable u_{it} has marginal distribution $N(0, \sigma^2)$ as in the LAR model.
It is worth noting that by imposing constraints on k, $\boldsymbol{\lambda}$, or $\boldsymbol{\Pi}$, special cases of the SW-LAR model result. In particular:

1. with $k = 1$ the LAR model described in Section 2 is obtained. The correlation coefficient is then the same for all subjects and occasions.

2. If the transition matrix is equal to an identity matrix, i.e. $\boldsymbol{\Pi} = \boldsymbol{I}$, the SW-LAR$_1$ model is obtained. Under this model, the correlation coefficient may be different between subjects belonging to different latent states, but not between occasions.

3. If the transition matrix has constant rows containing the initial probabilities, i.e. $\boldsymbol{\Pi} = \mathbf{1} \otimes \boldsymbol{\lambda}'$, the SW-LAR$_2$ model results, under which the correlation coefficient may change between subjects and occasions, since each subject randomly moves between different regimes.

3.2 Model estimation

We estimate the model parameters by maximizing the corresponding log-likelihood which is given by

$$\ell(\boldsymbol{\theta}) = \sum_i \log p(\boldsymbol{y}_i|\boldsymbol{X}_i),$$

where $\boldsymbol{\theta}$ is a short-hand notation for all model parameters, $\boldsymbol{y}_i$ is the response vector with elements y_{it}, $t = 1, \ldots, T$, and $\boldsymbol{X}_i$ is the corresponding matrix of covariates made of the vectors $\boldsymbol{x}_{it}$.
A crucial point is how to compute the *manifest probability* or *density* $p(\boldsymbol{y}_i|\boldsymbol{X}_i)$, which is based on a T-dimensional integral. For this aim we implemented an algorithm which is related to the sequential numerical integration method of Heiss (2008).
Let $q_{it}(u, v) = p(u_{it} = u, v_{it} = v, y_{i1}, \ldots, y_{it})$ and note that, for $t > 1$, this probability may be expressed as

$$q_{it}(u, v) = f(y_{it}|u) \sum_{v_0} \pi_{v_0 v} \int_{\Re} q_{i,t-1}(u_0, v_0) g(u|u_0, v) du_0,$$

with

$$q_{i1}(u, v) = p(y_{i1}|u) \rho_v g(u),$$

where $f(y_{it}|u) = p(y_{it}|u_{it} = u)$, $g(u)$ denotes the density function for the distribution of u_{i1}, and $g(u|u_0, v)$ denotes that for the distribution of u_{it} given $u_{i,t-1} = u_0$ and $v_{it} = v$, with $t > 1$. The algorithm we implemented is based on computing first $q_{i1}(u, v)$ and then $q_{it}(u, v)$ for $t = 2, \ldots, T$ by a suitable Gaussian quadrature. These probabilities are computed for u equal to every node of the quadrature and $v = 1, \ldots, k$. At the end, we obtain

$$p(\boldsymbol{y}_i|\boldsymbol{X}_i) = \sum_v \int_{\Re} q_{iT}(u, v) du,$$

again computed by a suitable quadrature. Note that, this algorithm closely resembles the recursive algorithm commonly used for the maximum likelihood estimation of hidden Markov models (Baum *et al.*, 1970).

TABLE 1. Results from the fitting of models LAR, SW-LAR$_1$, and SW-LAR$_2$.

	LAR	SW-LAR$_1$	SW-LAR$_2$
μ_1	7.3270	9.1515	7.6452
μ_2	4.1949	5.2750	4.3014
μ_3	1.0229	1.2479	0.9076
μ_4	-2.3763	-3.0282	-2.6919
female	-0.0572	0.0440	-0.0591
non white	-1.8515	-2.2072	-1.8758
education	1.5882	1.9401	1.6746
age	-0.1012	-0.1207	-0.0929
σ	2.9159	3.9973	3.2414
ρ_1	0.9550	0.4889	0.4414
ρ_2	–	0.9758	1.0000
λ_1	1.0000	0.2406	0.1268
λ_2	–	0.7594	0.8732
log-likelihood	-8884.7	-8795.6	-8818.2
# parameters	10	12	12
BIC	17838	17674	17719

In order to maximize $\ell(\boldsymbol{\theta})$ we implemented a numerical algorithm which, for the moment, may be only used to deal with LAR, SW-LAR$_1$, and SW-LAR$_2$ models. Future research will be devoted to the implementation of an algorithm to estimate the more general SW-LAR model and to obtain standard errors for the parameter estimates.

4 Application

The data used for the illustrative application come from the Health and Retirement Study conducted by the University of Michigan (for a detailed description see `http://www.rand.org/labor/aging/dataprod`). In particular, we considered a set of 1000 American people who self-evaluated their health status over 8 occasions. The health status is measured on a scale based on five grades: poor, fair, good, very good, and excellent. For each subject, some covariates are available: *gender*, *race*, *education*, and *age at each occasion of interview*.

We fitted three different models on these data: LAR, SW-LAR$_1$ ($\boldsymbol{\Pi} = \boldsymbol{I}$) with $k = 2$ latent states, and SW-LAR$_2$ ($\boldsymbol{\Pi} = \mathbf{1} \otimes \boldsymbol{\lambda}'$) with the same number of latent states. The main results are given in Table 1.

We note that the SW-LAR$_1$ model has the highest log-likelihood and the smallest value of the BIC index (Schwarz, 1978). With the inclusion of only two more parameters, this model shows a much better fit than the LAR model. The estimates of the two correlation coefficients under this model

are rather different; in particular we have an estimate equal to 0.49 for the 24% of subjects and equal to 0.98 for the remaining 76% of subjects. These two different levels of correlation correspond to two different levels of persistence of the effect of unobservable factors on the response variables.

Acknowledgments

F. Bartolucci and F. Pennoni acknowledge the financial support from the "Einaudi Institute for Economics and Finance" (Rome - IT) and from PRIN 2007.

References

Bartolucci, F., and Farcomeni, A. (2009). A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. *Journal of the American Statistical Association*, **104**, 816-831.

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2010). An overview of latent Markov models for longitudinal categorical data. *Technical report* `http://arxiv.org/abs/1003.2804`.

Baum, L.E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, **41**, 164-171.

Chi, E.M., and Reinsel, G.C. (1989). Models for longitudinal data with random effects and AR(1) errors. *Journal of the American Statistical Association*, **84**, 452-459.

Greene, W.H., and Hensher, D.A. (2003). A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Research, Part B*, **37**, 681-698.

Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, **57**, 357-384.

Heiss, F. (2008). Sequential numerical integration in nonlinear state space models for microeconometric panel data. *Journal of Applied Econometrics*, **23**, 373-389.

Lindsay, B., Clogg, C.C., and Greco, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, **86**, 96-107.

Schwarz, G.E. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

Wiggins, L.M. (1973). *Panel Analysis: Latent probability models for attitude and behaviour processes*. Amsterdam: Elsevier.

An Empirical Bayes Approach to Variable Selection and QTL Analysis

Haim Bar[1,2], James Booth[1], Martin T. Wells[1]

[1] Department of Statistical Science, Cornell University, Ithaca, NY 14853; <hyb2@cornell.edu>, <jb383@cornell.edu>, <mtw1@cornell.edu> .
[2] Communicating Author.

Abstract: We develop a model-based empirical Bayes approach to variable selection problems in which there are a large number of candidate predictors. A key assumption in our approach is that only a small fraction of the candidates are associated with the response variable. The method is motivated by modern applications such as quantitative trait loci (QTL) association studies.

Keywords: Variable selection; QTL analysis; Empirical Bayes; EM algorithm.

1 Introduction

We address the problem of variable selection in normal linear regression models when there are a large number of candidate explanatory variables, most of which have little or no effect on the dependent variable. We illustrate our method using a famous data set and compare our results with others in the literature.

1.1 Motivation

Automated methods for variable selection in normal linear regression models have long been studied in the literature. Recent work on this topic includes George and McCulloch (1993), Breiman (1995), and Casella and Moreno (2006). Virtually every statistical package contains an implementation of standard stepwise methods. Traditional regression problems typically involve a small number of explanatory variables and an analyst can make educated decisions as to which ones should be included in the regression model, and which should not. However, the new age of high speed computing and recent analytic needs and technological advances in genetics, for example, have dramatically changed this paradigm. It is common practice to use linear regression models to estimate the effects of hundreds or even thousands of predictors on a given response. These modern applications present major challenges. First, there is the so-called "large p, small n" problem, since the number of predictors (e.g. genetic markers in

a QTL study) often greatly exceeds the sample size. Methods controlling the experiment-wise false discovery rate in one predictor at a time analyses often result in few or no discoveries. Second, the model space is huge. For example, for a modest QTL study with 1000 markers, there are 2^{1000} possible models. This renders traditional search-based algorithms impractical. In this paper we propose an empirical Bayes, model-based approach to variable selection which we implement via a fast EM algorithm.

1.2 The Ozone Data

We demonstrate our method using the well-known air-pollution data set, first introduced by Breiman and Friedman (1985) to illustrate the ACE procedure. It consists of daily measurements of ozone concentration levels in the Los Angeles basin, collected over 330 days in 1976. There are eight meteorological explanatory variables, labeled by Friedman and Silverman (1989) and subsequent authors (e.g., Hastie and Tibshirani, 1990) by x_1 = Vandenburg 500 millibar height, x_2 = humidity, x_3 = inversion base temperature, x_4 = Sandburg Air Force Base temperature, x_5 = inversion base height, x_6 = Daggot pressure gradient, x_7 = wind speed, and x_8 = visibility. We refer mostly to the analysis in Lee, Nelder and Pawitan (2006, subsection 2.4.4) which also uses x_9, the day of the year. Selecting a first-order linear regression model can be done easily by checking all 2^9 possible models, but this strategy is not feasible when we wish to include second, or third order terms (with 2^{54} and 2^{219} possible models, respectively.)

2 A Statistical Model for Automatic Variable Selection

Denote the (numeric) responses by $y_i, i = 1, ..., N$. Suppose that for each response we have P measurements , $x_{ij}, j = 1, ..., P$, of covariates of interest (e.g. sex, population, age) which we want to include in the regression model. Suppose also that there are M putative variables $z_{ik}, k = 1, ..., M$, of which only a small subset should be included in the model. We assume that the response, y_i, can be modeled using an additive combination of the covariates:

$$y_i = \sum_{j=1}^{P} x_{ij}\beta_j + \sum_{k=1}^{M} z_{ik}\gamma_k u_k + \epsilon_i, \tag{1}$$

where β_j are fixed effects, $u_k \overset{iid}{\sim} N(\psi, \sigma_u^2)$, and $\gamma_k \overset{iid}{\sim} Ber(p)$ are indicator variables, taking the value 1 if the k^{th} putative variable $\mathbf{z}_k$, is included in the model, and 0 otherwise. In this context, the problem of variable selection can therefore be seen as an estimation procedure, where our main

interest is in the latent variables $\{\gamma_k\}$. Note that (1) can be expressed in the familiar matrix form, as

$$Y = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

where

$$\begin{aligned}
&\mathbf{X} = (x_{ij}),\\
&\beta = (\beta_1, ..., \beta_P)^\top,\\
&\mathbf{Z} = (z_{ik}\gamma_k),\\
&\mathbf{u} \sim N(\mathbf{1}_M\psi, \sigma_u^2\mathbf{I}_M),\\
&\mathbf{e} \sim N(\mathbf{0}_N, \sigma_e^2\mathbf{I}_N).
\end{aligned}$$

We employ an empirical Bayes approach in which the parameters β, ψ, σ_e^2, and σ_u^2 are estimated via a modified EM algorithm, and upon convergence, we select $\mathbf{z}_k$ to be included in the model if the estimated posterior probability of its latent indicator, γ_k, is greater than a predefined threshold. The complete data likelihood, $f_C(y, \gamma)$, is obtained by integrating out the random effects, $\{u_k\}$. Then the Q-function for the EM algorithm is given by $Q(\theta|\theta^{(t)}) = E_{\theta^{(t)}}\{\log f_C(y, \gamma)|y\}$.

Application of the EM algorithm is not entirely straightforward, for two reasons. First, the log complete data likelihood is a non-linear function of the latent variables, making the E-step analytically intractable. We solve this problem by updating the γ_k's by their posterior expectations using Bayes Rule. A second problem stems from the modeling of the putative variables as random effects. When we integrate out the random effect, the variance-covariance matrix of the posterior likelihood contains a large $(N \times N)$ matrix of the form $\mathbf{I}_N + \frac{\sigma_u^2}{\sigma_e^2}\mathbf{Z}\mathbf{Z}^\top$, which has to be inverted to compute the iterative maximum likelihood estimates. To address this computational problem we use the Woodbury identity (Golub and Van Loan, 1996) and express $f_C(y, \gamma)$ in terms of the $M \times M$ matrix $\Sigma_M^* = \mathbf{I}_M + \frac{\sigma_u^2}{\sigma_e^2}\mathbf{Z}^\top\mathbf{Z}$. This simplifies the computation because the $(k, l)th$ element of $\mathbf{Z}^\top\mathbf{Z}$ is given by $\langle \mathbf{z}_k, \mathbf{z}_l\rangle\gamma_k\gamma_l$, where $\langle , \rangle$ denotes the inner product of two vectors. In contrast, the elements of $\mathbf{Z}\mathbf{Z}^\top$ involve all the $\gamma_k s$. We propose to set $\gamma_k^{(t)} = 0$ if the posterior expectation of the k^{th} latent variable in the t^{th} iteration is below a given threshold. Since only a small number of the putative variables are truly associated with the response, the matrix Σ_M^* is relatively sparse and much easier to invert.

3 Results and Discussion

We have performed simulations involving thousands of putative variables, and achieved excellent results in terms of selecting the correct model, and speed. Here, we focus on the analysis of the ozone data.

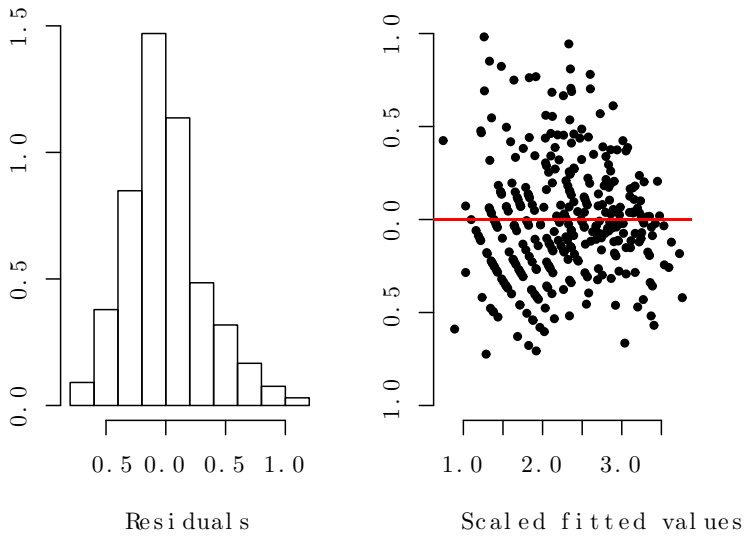


FIGURE 1. Ozone data – residuals diagnostic plots for model (2).

3.1 Case Study – The Ozone Data

Our model-based approach and the computational strategy that we described in Section 2 allow us to analyze a very large number of explanatory variables very efficiently. As an example, we applied our model selection approach on the ozone data with all second and third order terms (a total of 219 variables) as candidate predictors. Convergence was achieved in a few seconds, and we compared the selected model with those discussed in a recently published book by Lee, Nelder and Pawitan (2006, subsection 2.4.4). In terms of the Akaike Information Criterion (AIC), Lee et al. reported that the best result (their model 2.8) was obtained with a log link gamma GLM with linear predictor: $x_2 + x_4 + x_7 + x_8 + x_9 + x_8^2 + x_9^2$. This model has an AIC score of 1743.3. Our algorithm chose the combination of predictors,

$$x_3 + x_4 + x_5 + x_6 + x_7 + x_9 + x_3x_7 + x_6^2 + x_9^2, \tag{2}$$

with the resulting gamma GLM having AIC equal to 1702. Figure 1 shows the distribution of the residuals, and a scatter-plot of the residuals by scaled (log) fitted values. These diagnostics plots provide further evidence for the adequacy of the model selected by our method.

3.2 Extensions and QTL Applications

Our model was developed with the goal of analyzing very large QTL data sets in order to detect loci that are associated with biological traits. We are currently working on three important extensions to the model.

First, since some putative variables may have a positive effect on the response while others have a negative effect, we believe that better-fitting models may be obtained by allowing u_k in (1) to follow the mixture distribution:

$$u_k \sim N(\psi, \sigma_u^2) \text{ with probability } p_1, \text{ or}$$
$$u_k \sim N(-\psi, \sigma_u^2) \text{ with probability } p_2.$$

The second extension involves adding interactions between fixed effects and SNPs. For example, if the quantitative trait is the Forced Expiratory Volume (FEV), it may be the case that the significant loci have a different effect on the response for heavy smokers, than for non-smokers. Furthermore, it may be the case that certain SNP×SNP interactions will have a negative effect, while others have a positive effect, depending on the chromosomes on which they lie. This requires a relatively simple modification to the model and the estimation procedure.

The third enhancement is to extend the model to the generalized linear model framework in order to deal with binomial and Poisson responses, as well as censored survival times using the artificial Poisson model as described by Whitehead (1980).

We developed the software initially as a prototype in R, and are currently working on writing it in C, which we believe will increase its speed significantly, and will enable handling a much larger number of SNPs.

3.3 Conclusions

Our simulations and case studies (one described in this paper) show that our model-based approach to variable selection provides excellent results in terms of accuracy and speed. It allows adding higher-order terms to regression models, and can be extended to a wider range of applications, involving a large number of covariates.

References

Breiman, L. (1995). Better Subset Regression Using Nonnegative Garrote. *Technometrics*, **37**, 373-384.

Breiman, L., Friedman, J. (1985). Estimating Optimal Transformations for Multiple Regression and Correlation. *Journal of the American Statistical Association*, **80**, 580-598.

Casella, G., Moreno, E. (2006). Objective Bayesian Variable Selection. *Journal of the American Statistical Association*, **101**, 157-167.

Friedman, J., Silverman, B. (1989). Flexible Parsimonious Smoothing and Additive Modeling. *Technometrics*, **31**, 3-21.

George, E.I., McCulloch, R.E. (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, **88**, 881-889.

Golub, G.H., Van Loan, C.F. (1996), *Matrix Computations.* Baltimore: The Johns Hopkins University Press.

Hastie, T., Tibshirani, R. (1990), *Generalized Additive Models.* New York: Chapman and Hall.

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

Whitehead, J. (1980). Fitting Cox's Regression Model to Survival Data Using GLIM. *Applied Statistics*, **29**, 268-275.

A Conditional Corregionalized Linear Model for Bioclimatic Classification

X. Barber[1], J. Morales[1], A. López-Quílez[2], A. Mayoral[1], A. Barber[3]

[1] Center of Operations Research, Miguel Hernandez University
[2] Dpto. Estadística e I.O. Universitat de València
[3] IDENTIA Institute

Abstract: A spatial multivariate hierarchical model is proposed to establish a bioclimatic classification based on different climatic indexes. This model consists on a conditional corregionalized linear model that allows us to easily specify the relationship among the considered indexes, and also to improve the univariate modelling.

Keywords: Bayesian spatial models; multivariate hierarchical models; bioclimatic indexes; conditional corregionalized linear model.

1 Introduction

Bioclimatology or Phytoclimatology are important sciences for the comprehension of the close relationship between climate and vegetation, and therefore, the plant landscape. Thus, the better knowledge of the interrelationship climate-vegetation we get, the better management of plant resources, landscape, and environment can we accurately develop.
The main aim of this work is to improve a former Bioclimatic classification of the Island of Cyprus proposed in Barber (1995), based on the current Worldwide Bioclimatic Classification System (Rivas-Martinez, 2004).

2 Data and Bioclimatic index

Available data consists on different measures from 59 meteorological stations distributed all over the island of Cyprus. Geographical UTM coordinates and altitude are also available.
From the information above, several bioclimatic indexes derive. Some of the most relevant for explaining biodiversity are the Ombrothermic Index (OI), the Continentality Index (CI) and the Thermicity Index (TI). The Ombrothermic Index (OI):

$$OI = (Pp/Tp) \times 10,$$

where Pp is the Annual Positive Precipitation, and Tp is the Annual Positive Temperature. The Continentality Index (CI) as an annual thermic interval:

$$CI = T_{max} - T_{min},$$

assessed in Celsius degrees, and provides the range between the average temperatures of the warmest (T_{max}) and coldest (T_{min}) months of the year. Finally, the Thermicity Index (TI) is defied as

$$TI = (T + m + M)10,$$

ten times the sum of T (Annual Mean Temperature) + M (Mean Daily Maximum Temperature) + m (Mean Daily Minimum Temperature).

3 The Model

The relationship among these indexes suggests that a multivariate approach is needed to obtain a more accurate spatial representation to provide a correct bioclimatic classification. In Barber (2009) showed that thermicity operates independently of the OI and CI for these data.
We propose a conditional corregionalized linear model (Banerjee et al., 2004) using the reparameterization of the variance proposed in Yan et al. (2007) to predict the CI and OI on the whole island, just by knowing its value at the 59 meteorological stations.
Therefore, proposed bivariate model for CI and $LOI = log(OI)$ is:

$$\begin{array}{ll} (I) & \left\{ \begin{array}{l} \mathbf{Y}_{LOI} \sim N_n\left(\mathbf{X}^T\beta_{LOI}, \mathbf{\Sigma}_{\mathbf{W}_{LOI}}\right) \\ \mathbf{Y}_{CI}|\mathbf{Y}_{LOI} \sim N_n\left(\mathbf{X}^T\beta_{CI|LOI} + \alpha_{CI|LOI}\mathbf{Y}_{LOI}, \mathbf{\Sigma}_{\mathbf{W}_{CI}}\right) \end{array} \right. \\ (II) & p(\beta_{LOI}, \beta_{CI|LOI}, \alpha_{CI|LOI}, \kappa_{CI}, \xi^2_{LOI}, \xi^2_{CI}, \theta_{LOI}, \theta_{CI}), \end{array}$$

and assuming independent priors for all the parameters:

$$\begin{array}{l} \pi(\beta_{LIO}) \propto 1 \ ; \ \pi(\beta_{IC|LIO}) \propto 1 \ ; \ \ \pi(\alpha_{IC|LIO}) \propto 1 \\ \pi(\phi_{LIO}) = \pi(\phi_{IC}) \sim Unif(1,0E-05; 5,5E-04) \\ \pi(\nu_{LIO}) = \pi(\nu_{IC}) \sim Unif(0,05; 1,95) \ ; \\ \pi(\kappa_{IC}) \sim Unif(0;1) \\ \pi(\xi_{LIO}) \propto Unif(0.001; \sqrt{b_{LIO}}) \ ; \ \pi(\xi_{IC}) \propto Unif(0.001; \sqrt{b_{IC}}), \end{array}$$

with variances and covariances matrix given by:

$$\begin{array}{l} \mathbf{\Sigma}_{\mathbf{W}_{LOI}} = \xi^2\mathbf{H}(\theta); \\ \mathbf{\Sigma}_{\mathbf{W}_{CI}} = \xi^2\left[(1-\kappa)\mathbf{H}(\theta) + \kappa\mathbf{I}\right], \end{array}$$

where $X = (1, \textit{Elevation})$, W is the vector of spatial random effects, H is a correlation matrix (assuming the Matern family for correlations) between spatial locations with isotropic correlation function ρ, and ν the parameter

controlling the smoothness. ξ^2_{LOI} is variance for the log(OI) and ξ^2_{CI} for the CI. The ratio κ is interpreted as the fraction for the total variation of CI or LOI contributed by the measurement error. The ϕ is the decay parameter. We define the range (R) as $1/\phi$ and the effective range (ER) as the distance at which the correlation has dropped to only 0.05.

4 Results and Conclusions

We obtain the inferences based on the posterior distribution, and the prediction of the indexes in the whole island (see Figure 1). The Bayesian framework provides interesting summaries as probabilities for any location (see Figure 2) of belonging to some specific Bioclimatic-subtype: the Spatial Distribution Function (SDF) (Barber, 2009). Note that the gradation on the border given by the SDF is higher in the multivariate models than in the univariate.

Acknowledgments: This research was supported by the Spanish Ministry of Education and Science, under Grant MTM2007-61554. We do acknowledge the Meteorological Service of Cyprus for the data provided to Antoni Barber during his granted staying in Cyprus for researching at the Department of Forests (Government of Cyprus).

References

Banerjee, S., Carlin, B., and Gelfand, A. (2004). it Hierarchical Modeling and Analysis for Spatial Data. Chapman-Hall/CRC.

Barber, A (1995). Bioclimatology and Potential Vegetation of the Island of Cyprus. Diploma Thesis. Cyprus Forestry College & Department of Forests. Government of Cyprus.

Barber, X. (2009). Modelos geoestadísticos para el estudio de índices bioclimáticos *(Geoestatistical models for the study of bioclimatic indexes)*. PhD Thesis. Universidad Miguel Hernández de Elche.

Rivas-Martinez, S. (2004). Worldwide bioclimatic classication system. www.globalbioclimatics.org.

Yan, J., Cowles, M., Wang, S., and Armstrong, M. (2007). Parallelizing MCMC for bayesian spatiotemporal geostatistical models. *Statistics and Computing, 17(4):323-335.*

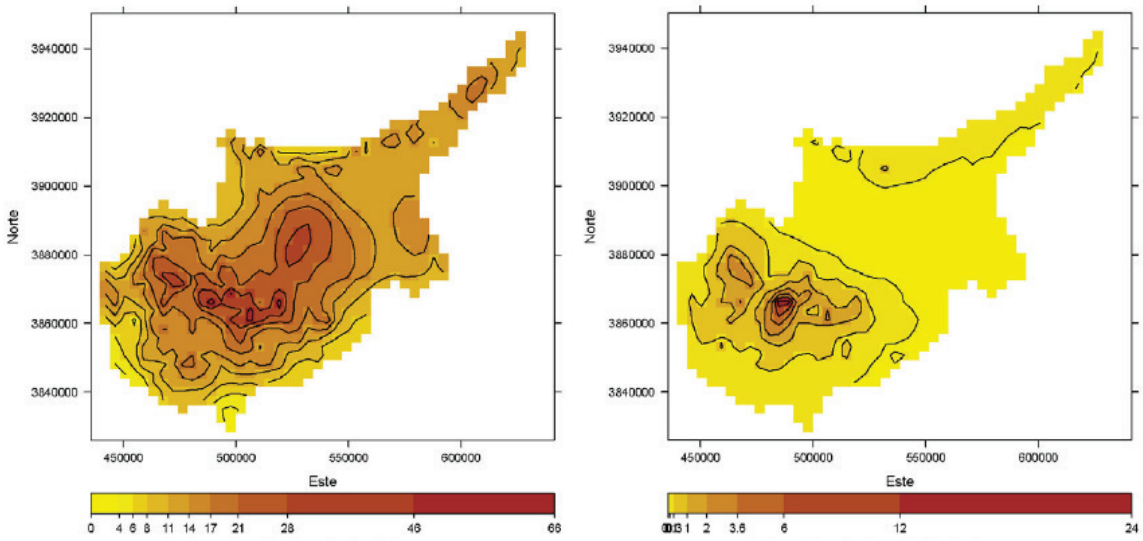


FIGURE 1. Prediction distribution: Continentality and Ombrothermic.

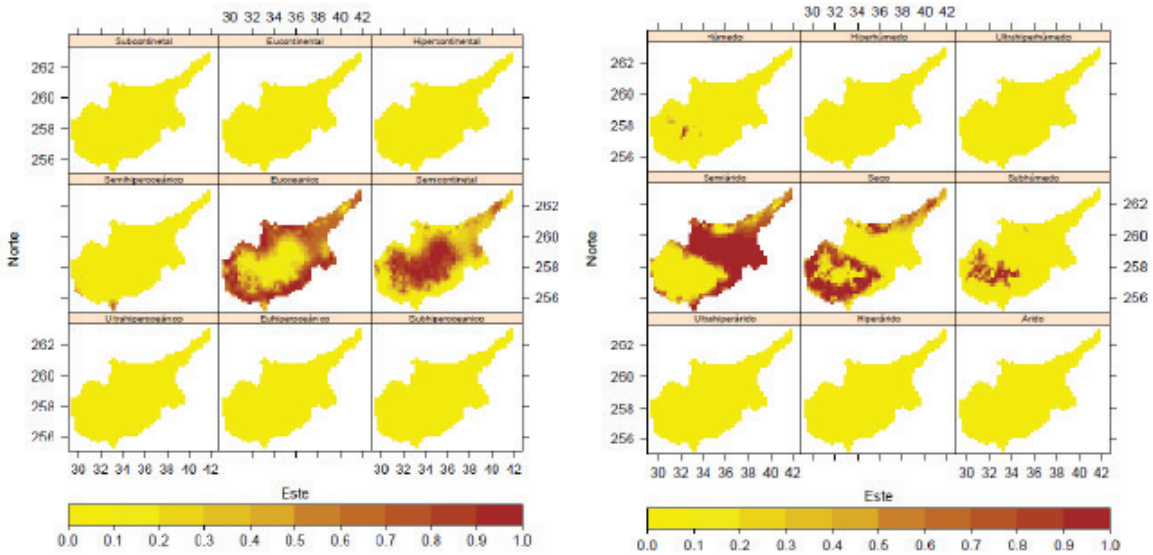


FIGURE 2. SDF: Continentality and Ombrothermic.

Time Series of Proportions: A Compositional Approach

C. Barceló-Vidal[1], L. Aguilar[2]

[1] Dept. Informàtica i Matemàtica Aplicada, Campus de Montilivi, Univ. de Girona, E-17071 Girona, Spain – carles.barcelo@udg.edu
[2] Dept. de Matemáticas, Escuela Politécnica, Univ. de Extremadura, E-10071 Cáceres, Spain – luciaaz@unex.es

Abstract: Taking account of the compositional nature of proportions, the logistic function is the natural transformation to apply when analyzing and modelling time series of continuous proportions. From the metric structure on the simplex, an isomorphic structure is defined on the set of continuous proportions. This structure permits the translation of the standard analysis of univariate time series to the compositional analysis of proportions in which the logistic transformation arises naturally and not as a mere alternative to the transformation of the data.

Keywords: Compositional data; log odds ratio; logistic transformation; simplex; time series of proportions.

1 Introduction

Univariate time series (TS) of proportions, p_t, arise in a wide variety of applications. Authors often ignore the restricted range of variation of the p_t, namely $(0, 1)$, and use standard techniques to model TS of proportions (e.g., Box and Jenkins, 1976; Tiller, 1992). Such analyses can result in estimates of proportions erroneously lying outside the interval $(0, 1)$.
Wallis (1987) was the first author to propose the logistic transformation $y_t = \operatorname{logit} p_t = \log(p_t/(1 - p_t))$ as an appropriate transformation for TS of proportions. The arguments given by him for the use of the logit transformation are: i) the necessity to stabilize the variance and make the transformed data approximately normally distributed, and ii) to ensure that estimates and projections lie within $(0, 1)$. Other authors use the logarithmic transformation as an alternative transformation to model TS of proportions because it can be useful as a means of stabilizing the variance and normalizing transformed data. However, it is not a guarantee that estimates and projections will lie in $(0, 1)$. It would appear then that the analysis and modelling of TS of proportions simply requires, if really necessary, the identification of the appropriate transformation in each case. However, in our opinion, this transformation based approach ignores the compositional nature of proportions. Any proportion p_t inevitably has associated

with it the complementary proportion $1 - p_t$, and thus the modelling of a TS of proportions p_t should be based upon the time series of compositions $\mathbf{p}_t = (p_t, 1 - p_t)'$ in the simplex $\mathcal{S}^2$.
This paper presents the compositional approach to the modelling of TS of proportions based on the initial work of Barceló-Vidal *et al.* (2007) which uses the compositional data analysis methodology introduced by Aitchison (1986) and subsequently developed by Egozcue *et al.* (2006). In Section 2 we present the Euclidean vector space $(¶, \oplus, \odot)$ of proportions in $(0, 1)$ in correspondence with the simplex $(\mathcal{S}^2, \oplus, \odot)$. The algebraic structure of ¶ serves as the basis of the development, in Section 3, of the compositional approach to the analysis of TS of proportions and to introduce the compositional ARIMA models. There it is shown that the logit transformation is the natural transformation which should be applied when attempting to analyze or model TS of proportions as it is the one that takes into account their compositional nature.

2 Continuous proportions as a compositional space

2.1 The metric space of continuous proportions

Let ¶ be the set of continuous proportions $p \in (0, 1)$. We identify a proportion p with the 2-part $\mathbf{p} = (p, 1 - p)' \in \mathcal{S}^2$ and therefore we can easily translate to ¶ the structure defined in $(\mathcal{S}^2, \oplus, \odot)$. The *perturbation* of p and p^* in ¶ will be denoted as

$$p \oplus p^* = \frac{pp^*}{pp^* + (1-p)(1-p^*)} = \frac{\operatorname{odds} p \times \operatorname{odds} p^*}{1 + \operatorname{odds} p \times \operatorname{odds} p^*} \ ,$$

where $\operatorname{odds} p = \frac{p}{1-p}$. The proportion $1/2$ is the neutral element of the group $(¶, \oplus)$, the *inverse* of p in $(¶, \oplus)$ is

$$\frac{\operatorname{odds} p}{1 + \operatorname{odds} p},$$

and the *compositional difference* between $p, p^* \in (0, 1)$ will be given by

$$p \ominus p^* = \frac{\frac{p}{p^*}}{\frac{p}{p^*} + \frac{1-p}{1-p^*}} = \frac{\frac{\operatorname{odds} p}{\operatorname{odds} p^*}}{1 + \frac{\operatorname{odds} p}{\operatorname{odds} p^*}} \ .$$

The *power transformation* of $p \in (0, 1)$ and $\alpha \in \mathbb{R}$ will be defined by

$$\alpha \odot p = \frac{p^\alpha}{p^\alpha + (1-p)^\alpha} = \frac{(\operatorname{odds} p)^\alpha}{1 + (\operatorname{odds} p)^\alpha} \ .$$

In this manner, $(¶, \oplus, \odot)$ becomes a one-dimensional real vector space. It is important to note that the algebraic structure of ¶ is based on the odds

of the proportions and thus not only takes into account of p but also its complement $1-p$.
The *additive logratio* transformation (alr) on $\mathcal{S}^2$ corresponds to the logit transformation on ¶, and the *centered* (or *symmetric*) *logratio* transformation (clr) corresponds to the $\frac{1}{2}$ logit transformation on ¶. As they are linear transformations from the vector space $(¶, \oplus, \odot)$ to $\mathbb{R}$ it holds that

$$\text{logit}\,((\alpha \odot p) \;\oplus\; (\alpha^* \odot p^*)) = \alpha \,\text{logit}\, p \;+\; \alpha^* \,\text{logit}\, p^*,$$

$$\text{logit}^{-1}\,(\alpha\, y \;+\; \alpha^*\, y^*) = (\alpha \odot \text{logit}^{-1} y) \;\oplus\; (\alpha^* \odot \text{logit}^{-1} y^*),$$

for any $p, p^* \in ¶$, and any $\alpha, \alpha^*, y, y^* \in \mathbb{R}$. Recall that the inverse of the logistic transformation can be expressed as $\text{logit}^{-1} y = \exp y / (1 + \exp y)$. It also holds that

$$\text{odds}\,((\alpha \odot p) \;\oplus\; (\alpha^* \odot p^*)) = (\text{odds}\, p)^{\alpha} \times (\text{odds}\, p^*)^{\alpha *}.$$

The $\mathcal{C}$-norm of a proportion $p \in (0,1)$ is given by

$$\|p\|_{\mathcal{C}} = \frac{1}{\sqrt{2}} \left|\text{logit}\, p\right|,$$

and the $\mathcal{C}$-distance between two proportions p and p^* in $(0,1)$ by

$$d_{\mathcal{C}}(p, p^*) = \|p \ominus p^*\|_{\mathcal{C}} = \frac{1}{\sqrt{2}} \left|\text{logit}\, p - \text{logit}\, p^*\right|.$$

The $\mathcal{C}$-norm converts the vector space $(¶, \oplus, \odot)$ into a metric space, and the $\frac{1}{2}$ logit transformation can be viewed as an isometry between ¶ and $\mathbb{R}$.

2.2 Compositional random continuous proportions

If p is a random continuous proportion in $(0,1)$, the *compositional* expected value ($\mathcal{C}$-mean) of p will be given by

$$\text{E}_{\mathcal{C}}\{p\} = \text{logit}^{-1}\left(\text{E}\left\{\text{logit}\, p\right\}\right).$$

In agreement with the concept of variance of a random variable and the $\mathcal{C}$-distance between two proportions, the *compositional* variance ($\mathcal{C}$-variance) of p will be defined as

$$\text{var}_{\mathcal{C}}\{p\} = \text{E}\left\{d^2_{\mathcal{C}}\left(p, \text{E}_{\mathcal{C}}\{p\}\right)\right\} = \text{E}\left\{\frac{1}{2}\left(\text{logit}\, p - \text{logit}\, \text{E}_{\mathcal{C}}\{p\}\right)^2\right\},$$

and, therefore, $\text{var}_{\mathcal{C}}\{p\} = \frac{1}{2}\text{var}\left\{\text{logit}\, p\right\}$.
Similarly, if (p, p^*) is a bivariate random proportion defined in $(0,1) \times (0,1)$, the *compositional* covariance ($\mathcal{C}$-covariance) and the *compositional* correlation ($\mathcal{C}$-correlation) of p and p^* will be defined as

$$\text{cov}_{\mathcal{C}}\{p, p^*\} = \tfrac{1}{2}\text{cov}\left\{\text{logit}\, p, \text{logit}\, p^*\right\},$$

$$\text{corr}_{\mathcal{C}}\{p, p^*\} = \text{corr}\left\{\text{logit}\, p, \text{logit}\, p^*\right\}.$$

The $\mathcal{C}$-mean and $\mathcal{C}$-variance of a random proportion p in $(0,1)$ are compatible with the algebraic structure of $(\P, \oplus, \odot)$ by which it holds that:

(i) $\mathrm{E}_{\mathcal{C}}\{p \oplus p^*\} = \mathrm{E}_{\mathcal{C}}\{p\} \oplus \mathrm{E}_{\mathcal{C}}\{p^*\}$;

(ii) $\mathrm{E}_{\mathcal{C}}\{\alpha \odot p\} = \alpha \odot \mathrm{E}_{\mathcal{C}}\{p\}$;

(iii) $\mathrm{var}_{\mathcal{C}}\{p \oplus p^*\} = \mathrm{var}_{\mathcal{C}}\{p\} + \mathrm{var}_{\mathcal{C}}\{p^*\} + 2\,\mathrm{cov}_{\mathcal{C}}\{p, p^*\}$;

(iv) $\mathrm{var}_{\mathcal{C}}\{\alpha \odot p\} = \alpha\,\mathrm{var}_{\mathcal{C}}\{p\}$,

for any $p, p^* \in \P$ and any $\alpha \in \mathbb{R}$.
Finally, the *compositional* normality of a random continuous proportion, p, will be associated with the normality of $\mathrm{logit}\, p$. Therefore, we will say that p is $\mathcal{C}$-normally distributed if $\mathrm{logit}\, p$ is normally distributed.
It would thus appear obvious that the compositional structure of the random continuous proportion p is based on that of the transformed proportion $\mathrm{logit}\, p$ and that the latter is compatible with the algebraic structure of $\P$ defined by the operators $\oplus$ and $\odot$.

3 Compositional approach to time series of proportions

From a *compositional* point of view, the time series analysis of continuous proportions p_t is based on the standard analysis of the series $\mathrm{logit}\, p_t$ and the fact that the algebraic operators on $\P$ that are compatible with this compositional approach are the perturbation operator $\oplus$ and the power transformation operator $\odot$, instead of the sum and multiplication by a scalar within in $\mathbb{R}$.

3.1 Some definitions

Let $p_t,\ t = 0, \pm 1, \pm 2, \ldots$ be a random process of continuous proportions in $(0,1)$. According to the compositional approach we define the $\mathcal{C}$-mean and the $\mathcal{C}$-variance of the process at time t as

$$\tilde{\mu}_t = \mathrm{E}_{\mathcal{C}}\{p_t\} = \mathrm{logit}^{-1}\left(\mathrm{E}\{\mathrm{logit}\, p_t\}\right); \quad \tilde{\sigma}_t^2 = \mathrm{var}_{\mathcal{C}}\{p_t\} = \tfrac{1}{2}\mathrm{var}\{\mathrm{logit}\, p_t\}.$$

Similarly, the $\mathcal{C}$-covariance and $\mathcal{C}$-correlation between p_{t_1} and p_{t_2} as

$$\tilde{\gamma}_{t_1,t_2} = \mathrm{cov}_{\mathcal{C}}\{p_{t_1}, p_{t_2}\} = \tfrac{1}{2}\mathrm{cov}\{\mathrm{logit}\, p_{t_1}, \mathrm{logit}\, p_{t_2}\},$$
$$\tilde{\varrho}_{t_1,t_2} = \varrho_{\mathcal{C}}\{p_{t_1}, p_{t_2}\} = \varrho\{\mathrm{logit}\, p_{t_1}, \mathrm{logit}\, p_{t_2}\}.$$

3.2 $\mathcal{C}$-stationarity and $\mathcal{C}$-white noise

A process of proportions p_t is called (weakly) $\mathcal{C}$-stationary if the following conditions are satisfied for all values of t:

$$\mathrm{E}_{\mathcal{C}}\{p_t\} = \tilde{\mu} = \text{constant}; \qquad \mathrm{cov}_{\mathcal{C}}\{p_t, p_{t+\tau}\} = \tilde{\gamma}(\tau),\ \tau = 0, \pm 1, \pm 2, \ldots$$

From a compositional perspective, a random process of proportions p_t is considered to be $\mathcal{C}$-*white noise* if

$$\mathrm{E}_{\mathcal{C}}\{p_t\} = 1/2, \quad \mathrm{var}_{\mathcal{C}}\{p_t\} = \tilde{\sigma}^2 \quad \text{and} \quad \mathrm{cov}_{\mathcal{C}}\{p_t, p_{t+\tau}\} = 0,$$

for $t = 0, \pm 1, \pm 2, \ldots$, and $\tau = \pm 1, \pm 2, \ldots$ Equivalently $\mathrm{logit}\, p_t$ should be white noise in the usual sense of the term, with variance $2\tilde{\sigma}^2$. We use the symbol ϵ_t to denote $\mathcal{C}$-white noise and represent by $\tilde{\sigma}_\epsilon^2$ the constant $\mathcal{C}$-variance of ϵ_t. If ϵ_t is $\mathcal{C}$-normally distributed, using the well known properties of the lognormal distribution, it is easy to prove that

$$\mathrm{E}\{\mathrm{odds}\, \epsilon_t\} = \exp\left(\tilde{\sigma}_\epsilon^2\right); \quad \mathrm{var}\{\mathrm{odds}\, \epsilon_t\} = \left(\exp(2\tilde{\sigma}_\epsilon^2) - 1\right)\exp(2\tilde{\sigma}_\epsilon^2).$$

3.3 The $\mathcal{C}$-difference operator

The $\mathcal{C}$-first difference operator $\nabla_{\mathcal{C}}$ is given by

$$\nabla_{\mathcal{C}} p_t = p_t \ominus p_{t-1} = (1 - L_{\mathcal{C}})p_t,$$

where $L_{\mathcal{C}}$ is the usual backshift operator. When the operator $L_{\mathcal{C}}$ is applied to a time series of proportions in a compositional context we have to take account of the algebraic structure of $(\P, \oplus, \odot)$. Thus, for example,

$$(1 - 2L_{\mathcal{C}} + L_{\mathcal{C}}^2)p_t = p_t \ominus (2 \odot p_{t-1}) \oplus p_{t-2}.$$

3.4 The $\mathcal{C}$-ARIMA model of proportions

A process of continuous proportions $p_t,\ t = 0, \pm 1, \pm 2, \ldots$, is a $\mathcal{C}$-ARMA(p, q) process if for every t,

$$p_t = (\phi_1 \odot p_{t-1}) \oplus \ldots \oplus (\phi_p \odot p_{t-p}) \oplus \epsilon_t \ominus (\theta_1 \odot \epsilon_{t-1}) \ominus \ldots \ominus (\theta_q \odot \epsilon_{t-q}), \quad (1)$$

where ϵ_t is $\mathcal{C}$-white noise $\mathcal{C}$-normally distributed with $\mathcal{C}$-variance $\tilde{\sigma}_\epsilon^2$. This equation can be written symbolically in the more compact form

$$\phi(L_{\mathcal{C}})(p_t) = \theta(L_{\mathcal{C}})\epsilon_t, \quad t = 0, \pm 1, \pm 2, \ldots,$$

where ϕ and θ are p^{th} and q^{th} degree polynomials in the $L_{\mathcal{C}}$ operator

$$\phi(L_{\mathcal{C}}) = 1 - \phi_1 L_{\mathcal{C}} - \ldots - \phi_p L_{\mathcal{C}}^p; \quad \theta(L_{\mathcal{C}}) = 1 - \theta_1 L_{\mathcal{C}} - \ldots - \theta_q L_{\mathcal{C}}^q.$$

Finally, a process of continuous proportions p_t is a $\mathcal{C}$-ARIMA(p,d,q) process if $(1-L_{\mathcal{C}})^d p_t$ is a $\mathcal{C}$-ARMA(p,q) process. It is clear that p_t is a $\mathcal{C}$-ARIMA(p,d,q) process if and only if $\operatorname{logit} p_t$ is a ARIMA(p,d,q) process. Therefore, in practice, the estimation of the parameters of a $\mathcal{C}$-ARIMA(p,d,q) process p_t reduces to the estimation of the parameters of the the transformed $\operatorname{logit} p_t$ process. $\mathcal{C}$-ARIMA(p,d,q) models can be represented in *logit* or *odds* formats. Thus, for example, equation (1) of a $\mathcal{C}$-ARMA(p,q) model can be expressed in *logit* format as

$$\begin{aligned}\operatorname{logit} p_t = \;& \phi_1 \operatorname{logit} p_{t-1} + \ldots + \phi_p \operatorname{logit} p_{t-p} \\ & + \operatorname{logit} \epsilon_t - \theta_1 \operatorname{logit} \epsilon_{t-1} - \ldots - \theta_q \operatorname{logit} \epsilon_{t-q},\end{aligned}$$

where $\operatorname{logit} \epsilon_t \sim \mathrm{N}(0, 2\tilde{\sigma}_\epsilon^2)$; and in *odds* format as

$$\operatorname{odds} p_t = (\operatorname{odds} p_{t-1})^{\phi_1} \times \ldots \times (\operatorname{odds} p_{t-p})^{\phi_p} \times \omega_t \times (\omega_{t-1})^{-\theta_1} \times \ldots \times (\omega_{t-q})^{-\theta_q},$$

where ω_t is log-normally distributed, i.e., $\omega_t \sim \Lambda(0, 2\tilde{\sigma}_\epsilon^2)$.

Acknowledgments: This research has been supported by the Spanish Ministry of Science and Innovation under the projects "CODA-RSS" (Ref. MTM2009-13272) and by the Agència de Gestió d'Ajuts Universitaris i de Recerca of the Generalitat de Catalunya (Ref. 2009SGR424).

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London, New York: Chapman & Hall. Reprinted in 2003 by Blackburn Press.

Barceló-Vidal, C., Aguilar, L., and Martín-Fernández, J.A. (2007). Compositional time series: a first approach. In: *Proceedings of the 22nd International Workshop of Statistical Modelling*, Barcelona, Spain, pp. 81–86.

Box, G.E.P., and Jenkins, G. (1976). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

Egozcue, J.J., and Pawlowsky-Glahn, V. (2006). Simplicial goemetry for compositional data. In: *Compositional Data Analysis: from Theory to Practice*, The Geological Society, London, UK, pp. 145-159.

Tiller, R.B. (1992). Time series modelling of sample data from the U.S. Current Population Survey. *Journal of Official Statistics*, **8**, 149–166.

Wallis, K.F. (1987). Time series analysis of bounded economic variables. *Journal Time Series Analysis*, **8**, 115–123.

Assessment of school performance through a multilevel latent Markov Rasch model

Francesco Bartolucci[1], Fulvia Pennoni[2], Giorgio Vittadini[3]

[1] Department of Economics, Finance, and Statistics, University of Perugia, Via A. Pascoli 20, 06123 Perugia, Italy, bart@stat.unipg.it
[2] Department of Statistics, University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy, fulvia.pennoni@unimib.it
[3] Department of Quantitative Methods for Business and Economic Sciences, University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy, giorgio.vittadini@unimib.it

Abstract: A multilevel version of the latent Markov Rasch model is proposed for the analysis of binary longitudinal data with covariates concerning pupils' proficiency. The model is formulated so that the distribution of the latent process, representing the evolution of the ability of each student, depends on random parameters related to the cluster to which the student belongs. Then, the effect of each cluster may be evaluated according to specific characteristics such as the type of school. Likelihood inference for the model, based on the EM algorithm, is outlined. An application based on data collected in an Italian Region is described.

Keywords: Binary longitudinal data; EM algorithm; Latent class model.

1 Introduction

The analysis of school performance, measured in terms of achievement attained by pupils at the end of each year of formal schooling, is usually based on value-added models; see Bryk and Weisberg (1976) for one of the first applications and, for a collection of papers dealing with these models, see the Spring 2004 issue of the *Journal of Educational and Behavioral Statistics*. One of the most used formulations of a value-added model assumes that the score of a student at the end of a school year is a linear function of the score at the end of the previous school year, individual covariates, and random effects depending on the school characteristics. The formulation of the distribution of these random effects reflects the multilevel structure of the data, due to students nested in classes.
In this paper, we propose an alternative approach for the analysis of school performance which is based on the latent Markov model (LM) of Wiggins (1973); for a review on this model and its extensions see Bartolucci *et al.* (2010). The LM model is a standard tool for the analysis of categorical longitudinal data when the interest is in describing individual changes with

respect to a certain latent status. The basic assumption of the model is that the response variables are conditionally independent given a latent process which follows a first-order Markov chain.
As outcomes we take the item responses (binary-scored) provided by each student at the end of each school year. Then, differently from the value-added approach, we do not need a univariate score for each student. Given the nature of the response variables, the model we propose is based on a Rasch (1961) type parametrization of the conditional distribution of these variables given the latent process; see also Bartolucci *et al.* (2008). In this way, the latent states are easily interpretable in terms of levels of ability. Moreover, the multilevel structure of the data is taken into account by allowing the initial and transition probabilities of the process for the ability to depend on a latent variable, which has the role of capturing the heterogeneity between clusters, further to time-constant and time-varying covariates. For this model we discuss maximum likelihood estimation based on the Expectation-Maximization (EM) algorithm (Baum *et al.*, 1970, Dempster *et al.*, 1977).
It is worth noting that the proposed approach is related to the mixture LM model (van de Pol and Langeheine, 1990, Kaplan, 2008) and to the LM model with random effects proposed by Altman (2007) in which, however, the random effects have a continuous distribution. The proposed model is also related to the multilevel latent transition model of Asparouhov and Muthén (2008) in which continuous latent variables are used to represent the effect of each cluster.
In the next section we outline the assumptions of the proposed model and its maximum likelihood estimation. Then, in Section 3 we describe the main results of the application to a dataset collected in an Italian Region and based on test scores on Mathematics administered over three years (from Grade 6 to 8) to students attending public and non-state middle-schools.

2 The proposed model

We consider a multilevel structure in which n subjects (students) are collected in C clusters (classes), with the c-th cluster having dimension n_c. Every subject is then identified by the pair of indices ci, with $c = 1, \ldots, C$ and $i = 1, \ldots, n_c$. We also denote by $\boldsymbol{Y}_{ci}^{(t)} = (Y_{ci1}^{(t)}, \ldots, Y_{cir_t}^{(t)})$ the vector of all r_t binary responses provided by this subject at occasion t, with $t = 1, \ldots, T$. Finally, we denote by $\boldsymbol{Y}_{ci}$ the set of the responses provided by this subject at all occasions and by $\boldsymbol{Y}_c$ the set of all responses provided by the subjects in the same cluster c.

2.1 Basic assumptions

The proposed multilevel extension of the LM model is based on the inclusion, for each cluster c, of a latent variable U_c having a discrete distribution

with support $\{1,\dots,k_1\}$. Moreover, as in a standard LM model, we represent the evolution of the ability of every subject ci by the sequence of latent variables $\boldsymbol{V}_{ci} = (V_{ci}^{(1)},\dots,V_{ci}^{(T)})$ which follows a first-order Markov chain with state space $\{1,\dots,k_2\}$ and parameters depending on U_c. We also assume that, for every cluster c, $\boldsymbol{V}_{c1},\dots,\boldsymbol{V}_{cn_c}$ are conditionally independent given U_c and that each response variable $Y_{cij}^{(t)}$ is conditionally independent of any other variable in the model given $V_{ci}^{(t)}$.
From the above assumptions it follows that, for every c, the response vectors $\boldsymbol{Y}_{ci},\dots,\boldsymbol{Y}_{cn_c}$ are conditionally independent given the latent variable U_c. In addition, the vector of response variables $\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_C$, associated to the different clusters, are marginally independent.
The model specification is completed by a Rasch (1961) parametrization of the distribution of the response variables given the latent process and by a suitable parametrization of the distribution of the latent variables. In particular, we assume

$$\lambda_j^{(t)}(v) = p(Y_{cij}^{(t)} = 1|V_{ci}^{(t)} = v) = \frac{\exp(\theta_v - \beta_j^{(t)})}{1+\exp(\theta_v - \beta_j^{(t)})},$$

with $t = 1,\dots,T$, $j = 1,\dots,r_t$, and $v = 1,\dots,k_2$, where θ_v is the ability level of the examinees in latent state v and $\beta_j^{(t)}$ is the difficulty level of the item. Moreover, about the initial probabilities $\pi_{ci}(v|u) = p(V_{ci}^{(1)} = v|U_c = u, \boldsymbol{z}_{ci}^{(1)})$ of the latent process for the ability, we assume

$$\log\frac{\pi_{ci}(v+1|u)+\cdots+\pi_{ci}(k_2|u)}{\pi_{ci}(1|u)+\cdots+\pi_{ci}(v|u)} = \delta_{0u} + \delta_{1v} + (\boldsymbol{z}_{ci}^{(1)})'\boldsymbol{\delta_2}, \tag{1}$$

with $u = 1,\dots,k_1$ and $v = 1,\dots,k_2-1$, where $\boldsymbol{z}_{ci}^{(t)}$ is a vector of covariates for subject ci at occasion t. Note that the intercepts δ_{0u} depend on the level of U_c; in order to ensure model identifiability, we set $\delta_{01} \equiv 0$.
Similarly, for what concerns the transition probabilities $\pi_{ci}^{(t)}(v_1|u,v_0) = p(V_{ci}^{(t)} = v_1|U_c = u, V_{ci}^{(t-1)} = v_0, \boldsymbol{z}_{ci}^{(t)})$, we assume

$$\log\frac{\pi_{ci}^{(t)}(v_1+1|u,v_0)+\cdots+\pi_{ci}^{(t)}(k_2|u,v_0)}{\pi_{ci}^{(t)}(1|u,v_0)+\cdots+\pi_{ci}^{(t)}(v_1|u,v_0)} = \eta_{0u}^{(t)} + \eta_{1v_0v_1}^{(t)} + (\boldsymbol{z}_{ci}^{(t)})'\boldsymbol{\eta}_2^{(t)}, \tag{2}$$

with $u = 1,\dots,k_1$, $v_0 = 1,\dots,k_2$, $v_1 = 1,\dots,k_2-1$, and $t = 2,\dots,T$, where we set $\eta_{01} \equiv 0$ in order to ensure model identifiability.
Finally, the conditional distribution of U_c given the covariates in $\boldsymbol{x}_c$ referred to the c-th cluster is modeled through a logit parametrization of the probabilities $\rho_c(u) = p(U_c = u|\boldsymbol{x}_c)$, that is

$$\log\frac{\rho_c(u+1)}{\rho_c(1)} = \gamma_{0u} + \boldsymbol{x}_c'\boldsymbol{\gamma}_{1u}, \quad u = 1,\dots,k_1-1. \tag{3}$$

2.2 Likelihood inference on the model parameters

The log-likelihood of the multilevel extension of the LM model may be expressed as

$$\ell(\boldsymbol{\phi}) = \sum_c \log p(\boldsymbol{Y}_c = \boldsymbol{y}_c),$$

where $\boldsymbol{\phi}$ is a short-hand notation for all model parameters and $\boldsymbol{y}_c$ is the set of observed responses for the subjects in the c-th cluster.
We maximize $\ell(\boldsymbol{\phi})$ by means of the EM algorithm (Baum *et al.*, 1970, Dempster *et al.*, 1977). As usual, this algorithm alternates two steps until convergence. At the E-step, the posterior distribution of each latent variable is computed. At the M-step, the expected value of the complete log-likelihood, computed on the basis of these posterior distributions, is maximized by standard tools. An important point is the initialization of the EM algorithm. Different strategies of initialization, such as the one based on random starting values, may be used in order to overcome the problem of multimodality of the likelihood.
Regarding the application of the model, a crucial aspect concerns the choice of the number of support points of the latent variables at cluster level (k_1) and the number of latent states for the ability (k_2). For this aim, we rely on the Bayesian Information Criterion (BIC) proposed by Schwarz (1978), which is based on the minimization of the index

$$BIC = -2\ell(\hat{\boldsymbol{\phi}}) + g\log(n),$$

where $\ell(\hat{\boldsymbol{\phi}})$ is the maximum log-likelihood of the model with given values of k_1 and k_2 and g is the corresponding number of parameters.
Finally, hypotheses on the model parameters can be tested by using the likelihood ratio statistic $-2[\ell(\hat{\boldsymbol{\phi}}_0) - \ell(\hat{\boldsymbol{\phi}})]$, where $\hat{\boldsymbol{\phi}}_0$ is the estimate of the parameter vector under the hypothesis of interest, which can be again computed through the EM algorithm. When the hypothesis is on one of the regression coefficients in (1), (2), or (3), we can alternatively use a Wald-test based on their standard errors. We obtain these standard errors on the basis of the numerical derivative of the score vector (first derivative of the log-likelihood with respect to $\boldsymbol{\theta}$).

3 Application

The application motivating this paper is part of a project to asses the level of students' achievement in Mathematics in Lombardy (Italy).
The data derive from the repeated administration of a set of dichotomously-scored items to a cohort of 1,246 students attending public and non-state middle-schools in the region. The items were administered at the end of each school year (28 at the end of the first year, 30 at the end of the second, and 39 at the end of the third). The individual covariates which

TABLE 1. Average class probabilities (for the latent variables V_c) stratified by type of school, years since school opened, and students/teachers ratio.

	A	B	C	D
≥ 17.5 *years*	0.425	0.189	0.314	0.072
< 17.5 *years*	0.363	0.196	0.289	0.152
public	0.325	0.233	0.376	0.066
non-state	0.786	0.000	0.000	0.214
≥ 8	0.541	0.090	0.205	0.164
< 8	0.337	0.245	0.363	0.054

are available in the sample are the *father and mother education*, which we decided to include separately, as they play a different role on the student's performance. The available covariates at cluster level are the following: *type of school*, *students/teachers ratio*, and *year since school opened*.

On these data we fitted the model illustrated in Section 2 with different values of k_1 and k_2. On the basis of the BIC index, we selected the model based on $k_1 = 4$ support points for the latent variables U_c and $k_2 = 6$ states for each latent process $\boldsymbol{V}_{ci}$. Every latent state v is associated to an estimate of the ability level θ_v, so that we can easily identify the group of the most proficient students and that of the least proficient students. Such levels may represent some specific types of task in Mathematics that a student is likely to perform successfully. The estimated conditional probabilities are higher for the items administered at the lower grades meaning that the difficulty of the items is increasing over time.

More interesting are the estimates of the parameters δ_{0u} and $\eta_{0u}^{(t)}$, which allow us to identify four different types of classes of students (A, B, C, D), according to the impact on the ability level. Classes of type A, B, and D have a similar effect on the ability at the first occasion, which is smaller than the effect of classes of type C. More heterogeneous is the effect on the evolution of the ability from the first to the third year. Overall, classes of type D show the best effect on the evolution of the ability, whereas the other classes have a similar overall effect.

On the basis of the estimates of the parameters $\boldsymbol{\gamma}_{1u}$ we can identify how the covariates related to each class affect the probability that this class is of type A, B, C, or D. According to Table 1 the most interesting conclusion is that classes of students in public schools are mostly of type A, B, or C, whereas those in non-state schools are only of type A or D.

Acknowledgments: F. Bartolucci and F. Pennoni acknowledge the financial support from the "Einaudi Institute for Economics and Finance" (Rome - IT) and from PRIN 2007.

References

Altman, R.M. (2007). Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. *Journal of the American Statistical Association*, **102**, 201-210.

Asparouhov, T., and Muthén, B. (2008). Multilevel mixture models. In: *Advances in latent variable mixture models*, G.R. Hancock and K.M. Samuelson (Eds.), 27-51, Charlotte, NC: Information Age Publishing.

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2010). An overview of latent Markov models for longitudinal categorical data. *Technical report* `http://arxiv.org/abs/1003.2804`.

Bartolucci, F., Pennoni, F., and Lupparelli, M. (2008). Likelihood inference for the latent Markov Rasch model. In: C. Huber, N. Limnios, M. Mesbah, and M. Nikulin (Eds.), *Mathematical Methods for Survival Analysis, Reliability and Quality of Life*, 239-254, Wiltshire: Wiley.

Baum, L.E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, **41**, 164-171.

Bryk, A.S., and Weisberg, H.I. (1976). Value-added analysis: a dynamic approach to the estimation of treatment effects. *Journal of Educational Statistics*, **1**, 127-155.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.

Kaplan, D. (2008). An overview of Markov chain methods for the study of stage-sequential developmental processes. *Developmental Psychology*, **44**, 457-467.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In: *Proceedings of the IV Berkeley Symposium on Mathematical Statistics and Probability*, **4**, 321-333.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

van de Pol, F., and Langeheine, R. (1990). Mixed Markov latent class models. *Sociological Methodology*, **20**, 213-247.

Wiggins, L.M. (1973). *Panel Analysis: Latent Probability Models for Attitude and Behavior Process*. Amsterdam: Elsevier.

Likelihood inference for a semi-parametric causal model addressing partial compliance by continuous principal strata

Francesco Bartolucci[1], Leonardo Grilli[2]

[1] Dipartimento di Economia, Finanza e Statistica - Università di Perugia, Via A.Pascoli, 20 - 06123 Perugia, bart@stat.unipg.it
[2] Dipartimento di Statistica - Università di Firenze, Viale Morgagni, 59 - 50134 Firenze, grilli@ds.unifi.it (COMMUNICATING AUTHOR)

Abstract: We propose a semi-parametric model for causal inference in the presence of continuous principal strata. The context of application is that of clinical trials with partial compliance in both placebo and drug arms. In order to reduce the modelling assumptions, we link the observed marginal distributions of drug and placebo compliances by a Plackett copula, so that their association is modelled by a single parameter, whereas no restrictions are posed on the marginal distributions. A likelihood inference approach for this model is proposed. In particular, the association parameter between the two compliances is studied via profile likelihood. We apply the model to data previously analysed by Efron and Feldman (1991) and Jin and Rubin (2008), comparing assumptions and results.

Keywords: Copula; EM algorithm; principal causal effects; profile likelihood; randomized clinical trials.

1 Principal stratification and partial compliance

Principal stratification (Frangakis and Rubin, 2002) is a general framework for counterfactual causal inference in the presence of a problematic post-treatment variable, i.e. a variable that in the causal path is located between the treatment and the outcome and cannot be ignored. An example is given by the degree of compliance, which may take the simple form of all-or-none compliance or the more complex form of partial compliance. Principal stratification is an effective framework for analysing data from trials with imperfect compliance: experiments with all-or-none compliance imply discrete principal strata, whereas partial compliance entails continuous principal strata. The second instance is considerably more complex and more demanding in terms of modelling assumptions. This is one of the reasons why principal stratification was not used for analysing trials with partial compliance, until the important work of Jin and Rubin (2008), hereafter JR, who re-analysed the data of the classical paper of Efron and Feldman (1991), hereafter EF.

The EF data derive from a placebo-controlled double-blinded randomised clinical trial designed to study the effectiveness of a certain drug for lowering cholesterol levels. The dataset concerns 335 men (164 assigned to active pills of the drug and 171 assigned to placebo pills), who were randomly assigned packets of drug or placebo. For every subject it is known the average decrease of cholesterol during the study and the compliance degree; the latter is computed as the average proportion of assigned packets which were not returned. For these data, partial compliance is problematic since, for patients assigned to placebo, the observed cholesterol reduction is stronger at higher levels of compliance. Then, the effect of the compliance to drug is likely to be a combination of a "genuine" effect, due to the amount of drug taken, and a "fictitious" effect, due to the correlation with some unobserved characteristics, such as the propensity to a healthy living style.
The analysis performed by EF relies on the assumption that drug and placebo compliances are linked by a deterministic function, which is used to impute the placebo compliance to patients assigned to drug and viceversa. This assumption was judged to be overly restrictive by JR, who cast the problem in a principal stratification framework and specify a stochastic relationship through a parametric model based on Beta distributions. This specification, however, entails *negative side-effect monotonicity*, i.e. the drug compliance is no larger than the placebo compliance.
The specification of a parametric model for the drug and placebo compliances is a critical point since the two compliances are not jointly observed. Then, the empirical evidence on their relationship is yielded by the data only indirectly through the model on the potential outcomes. The scarcity of the empirical support on this relationship, however, may be masked in a fully parametric Bayesian analysis, such as the one of JR, with results overly sensitive to the model specification. Moreover, the assumption of negative side-effect monotonicity might be unduly restrictive.
To investigate the above issues, we propose a model in which the marginal distributions of the compliances to placebo and drug are left unconstrained, whereas their joint distribution is formulated through the Plackett copula (Nelsen, 2006), avoiding monotonicity assumptions. This distribution thus depends on a single association parameter which can be studied via profile likelihood and allows us to make a straightforward sensitivity analysis to investigate the role of this parameter in determining the inferential results. Specifically, we first estimate the marginal distributions of the compliances via the empirical distribution function and then, for each given value of the association parameter, we maximize the likelihood through an EM algorithm (Dempster *et al.*, 1977). In the application to the EF data, the approach turns out to be flexible and easy to implement, yielding an alternative and appealing way of modeling and interpreting such data.
In the following, we first illustrate the main assumptions of the proposed causal model and the related likelihood inference approach. Then, we summarise the results of the application of this model to the EF data.

2 The proposed model

For any individual i, Z_i is a binary variable equal to 1 if the subject was assigned to the treatment arm (drug) and equal to 0 if he was assigned to the control arm (placebo). The potential compliances are represented by the couple (d_i, D_i), where d_i denotes the compliance to placebo (observed if $Z_i = 0$ and missing if $Z_i = 1$), and D_i denotes the compliance to drug (observed if $Z_i = 1$ and missing if $Z_i = 0$). Both d_i and D_i are proportions and thus lie on the unit interval. The outcome variable has two potential versions denoted by $Y_i^{(0)}$ (under placebo) and $Y_i^{(1)}$ (under drug).
As in JR, we adopt a principal stratification approach, where principal strata are defined by the values of the couple (d_i, D_i). The *Principal Causal Effect* in stratum (d_i, D_i) is then $PCE(d_i, D_i) = E(Y_i^{(1)} - Y_i^{(0)} \mid d_i, D_i)$.
In formulating the distribution of the potential outcomes, we follow EF and JR and assume that, given (d_i, D_i), both $Y_i^{(0)}$ and $Y_i^{(1)}$ have normal distributions, even if we allow more flexible specifications for both the mean (including interactions) and the variance (allowing for heteroscedasticity). However, the core of our approach is the semi-parametric specification of the joint distribution of the compliances through a copula, i.e. a family of functions that can be applied to the marginal distributions of two random variables to obtain a valid joint distribution. In our application, we estimate the marginal distributions of d_i and D_i in a non-parametric way through the empirical distribution function and we model their association by the Plackett copula (Nelsen, 2006), which is characterized by a single parameter ψ that regulates the degree of association.
To perform maximum likelihood (ML) estimation of the model parameters we devise a procedure with the following steps: (*i*) compute the empirical distribution functions for d_i and D_i; (*ii*) fix the Plackett parameter ψ to a value taken from a suitable grid; (*iii*) for each given ψ and on the basis of these distribution functions, estimate the parameters of the regression models for the potential outcomes using an EM algorithm; and (iv) make inference on ψ through the profile likelihood.

3 Results of the application on the EF data

After a standard selection procedure, we choose a model where the regression equations for the potential outcomes have heteroscedasticity and include an interaction term $d_i D_i$ between the two compliances, whereas the estimate of the Plackett association parameter ψ is 17.727. The latter corresponds to a Pearson correlation of 0.689 between d_i and D_i.
Inference under the selected model needs some care since the estimates of the regression coefficients depend on the association parameter ψ, for which we have little empirical support. In fact, although the profile likelihood for ψ has a single maximum, it is rather flat. Thus we warn against only relying

on the inference derived from the point estimate of ψ. At this estimate, the joint distribution of the drug and placebo compliances is notably different from JR's. This can be appreciated by comparing scatter plots of random draws. The plot of JR shows an accumulation of points along the bisectrix ($D_i = d_i$), which suggests that their monotonicity assumption ($D_i \leq d_i$) may be too restrictive. Indeed, 21.6% of the points in our plot goes beyond the bisectrix, corresponding to individuals with positive side effects ($D_i > d_i$), even if in most cases D_i is only slightly larger than d_i. At the ML estimate of ψ, the estimated principal causal effects are $PCE(d_i, D_i) = (-21.878 + 73.359 d_i) D_i$. Then, differently from JR, the dependence of the PCE on the dose of the taken treatment is stronger at higher levels of placebo compliance due to the interaction term.
The fragile identification of ψ prompts us to perform a sensitivity analysis to assess how the estimated PCE depends on the value of this association parameter. Bounds for PCE should always be reported when they are wide or when the profile likelihood is rather flat or multimodal. To compute these bounds it is advisable to choose a set of values of ψ corresponding to the profile likelihood beyond a certain threshold; for example, we take the set of values for which the hypothesis of independence ($H_0 : \psi = 1$) is rejected. It is comforting to note a stable PCE at least for compliance values (d_i, D_i) around the observed medians, whereas wide intervals at "unusual" compliance levels are not necessarily cause for concern. In particular, the PCE at the median point (0.89, 0.70) is quite stable, ranging from 27.4 to 34.8. If we consider points with D_i larger than d_i, such as $d_i = 0.59$ (1st quartile) and $D_i = 0.95$ (3rd quartile), the PCE ranges from 14.0 to 29.5, but there is still evidence of a positive effect.

References

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, series B*, **39**, 1–38.

Efron, B., and Feldman, D. (1991). Compliance as an Explanatory Variable in Clinical Trials. *Journal of the American Statistical Association*, **86**, 9–17.

Frangakis, C. E., and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, **58**, 21–29.

Jin, H., and Rubin, D. B. (2008). Principal Stratification for Causal Inference With Extended Partial Compliance. *Journal of the American Statistical Association*, **103**, 101–111.

Nelsen, R.B. (2006). *An introduction to Copulas (2nd edition)*. New York: Springer.

Joint models for classification and comparison of mortality in different countries.

Viani D. Biatat[1], Iain D. Currie[1]

[1] Department of Actuarial Mathematics and Statistics, and the Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh, EH14 4AS.

Abstract: We propose a class of additive generalized linear array models (GLAMs) which facilitate the classification and comparison of mortality tables. Different mortality tables are modelled in terms of their distances (gaps) from a reference table. These gaps are smooth functions of age and/or time and provide a simple graphical summary of the differences between tables. In the paper we describe the models, discuss their computational demands and their resolution with GLAM. We present the results, largely graphical, of applying our methods to various mortality tables taken from the Human Mortality Database and from the Continuous Mortality Investigation Bureau.

Keywords: Mortality classification, dispersion, joint modelling, P-splines, GLAM.

1 Introduction

We suppose that we have mortality data for p populations, $p \geq 2$, consisting of death counts and exposures, arranged in $n_a \times n_t$ matrices $\boldsymbol{D}^{[r]}$ and $\boldsymbol{E}^{[r]}$, $r = 1, ..., p$, such that the rows and columns of $\boldsymbol{D}^{[r]}$ and $\boldsymbol{E}^{[r]}$ are classified respectively by ages $\boldsymbol{x}_a$ and years $\boldsymbol{x}_t$, each arranged in ascending order; their vector equivalents will be denoted by $\boldsymbol{d}^{[r]} = vec(\boldsymbol{D}^{[r]})$ and $\boldsymbol{e}^{[r]} = vec(\boldsymbol{E}^{[r]})$. For a single population, it is common and natural to suppose that there is a 2-dimensional smooth surface that drives the force of mortality. However, mortality data for two (or more) populations can have some connections between them. Two typical examples are (a) mortality for females and males where the latter is known to be heavier than that of the former, and (b) mortality by lives and by amounts (in life insurance) where the latter is known to be lighter than that of the former. In addition to that, male and female mortality (for example) generally have some similarities in their dynamism. In general, how much can the dynamism of p mortality tables be similar/different? Can we build a joint and economical model for mortality tables which are similar (in some way)? In this paper, we propose a class of additive models with different components for the economical modelling and comparison of such mortality tables: the first component describes a (common) two-dimensional smooth surface (viewed as the reference) and the remaining components describe the relative differences (gaps) between

these tables. This class of models leads to the classification of populations into different categories.

2 Model specifications

In population r, $r = 1, ..., p$, we suppose that the number of deaths $D_{i,j}^{[r]}$ at age i in year j can be described approximately by the over-dispersed Poisson assumption with mean $E_{i,j}^{[r]} \times \mu_{i,j}^{[r]}$, where $\mu_{i,j}^{[r]}$ is the force of mortality; we assume that the Poisson variance in population r is inflated by some positive factor ϕ_r: $var(D_{i,j}^{[r]}) = \phi_r \times E_{i,j}^{[r]} \times \mu_{i,j}^{[r]}$, where the ϕ_r's are the dispersion parameters.

In general, our models apply to any number of populations but, for simplicity, we present the work for two populations (1 and 2), with some discussion in the general situation of p populations. The key idea is the following: if the dynamism of the two populations is similar, then the relative variation of their forces of mortality can be captured by a moderate number of parameters, ie, if we set (conceptually) a 2-dimensional smooth surface for the force of mortality in population 1 (viewed as the reference), then the smooth force of mortality for population 2 can be captured by adding a "simple" gap to this reference. We describe two populations as *very similar* if the gap (relative variation) between them is constant in age and time; they would be *similar in time/age* if the gap is smooth (flexible) in age/time and constant in time/age; we would say that they are *similar* if the gap is additively smooth in both age and time; otherwise, they are *different*. Note that *very similar* populations are nested within *similar in time/age* populations, and *similar in time/age* populations are in turn nested within *similar* populations; hence for space reasons, only the model for *similar* populations will be detailed in this paper with some discussions and illustrations for the other two scenarios.

The first component (reference) of our models uses 2-dimensional P-splines (Eilers and Marx, 1996, Currie et al., 2004). Let $\boldsymbol{B}_a$, $n_a \times c_a$, and $\boldsymbol{B}_t$, $n_t \times c_t$, be the marginal regression matrices (which are 1-dimensional regression matrices of B-splines evaluated along age ($\boldsymbol{x}_a$) and year ($\boldsymbol{x}_t$) respectively); the Kronecker product $\boldsymbol{B}_t \otimes \boldsymbol{B}_a$ creates a 2-dimensional regression basis. If we denote by $\boldsymbol{y}^{[r]} = \boldsymbol{d}^{[r]}/\boldsymbol{e}^{[r]}$, the vector of observed forces of mortality in population r, then taking population 1 as the reference, the linear predictor of its force of mortality can be expressed as

$$\log\left(\mathbb{E}\left[\boldsymbol{y}^{[1]}\right]\right) = (\boldsymbol{B}_t \otimes \boldsymbol{B}_a)\,\boldsymbol{\theta}^{[1]}. \tag{1}$$

We use a rich basis of B-splines for age and year; a smooth surface is then obtained by marginal penalization; ie the coefficient vector $\boldsymbol{\theta}^{[1]}$ is subject

to the penalty

$$\boldsymbol{P}^{[1]} = \lambda_a \boldsymbol{I}_{c_t} \otimes \boldsymbol{\Delta}_a' \boldsymbol{\Delta}_a + \lambda_t \boldsymbol{\Delta}_t' \boldsymbol{\Delta}_t \otimes \boldsymbol{I}_{c_a}, \tag{2}$$

where $\boldsymbol{\Delta}_a$ and $\boldsymbol{\Delta}_t$ are second order difference matrices (of appropriate size), λ_a and λ_t are smoothing parameters in the age and year direction, and $\boldsymbol{I}_n$ is the identity matrix of size n. With this setting, if we assume that population 2 is *similar* to population 1, then we express the linear predictor of population 2 as:

$$\log\left(\mathbb{E}\left[\boldsymbol{y}^{[2]}\right]\right) = (\boldsymbol{B}_t \otimes \boldsymbol{B}_a)\,\boldsymbol{\theta}^{[1]} + (\mathbf{1}_{n_t} \otimes \boldsymbol{B}_a)\,\boldsymbol{\theta}^{[2,1]} + (\boldsymbol{B}_t \otimes \mathbf{1}_{n_a})\,\boldsymbol{\theta}^{[2,2]}, \tag{3}$$

where $\mathbf{1}_n$ is the n length vector of ones, and $\boldsymbol{\theta}^{[2,1]}$ and $\boldsymbol{\theta}^{[2,2]}$ are coefficient vectors quantifying the gaps. In (3), we require the second term in the right hand side to capture both the constant component and the smooth age dependent component of the gap, while the third term models only the smooth year dependent component of the gap. Hence we smooth $\boldsymbol{\theta}^{[2,1]}$ and $\boldsymbol{\theta}^{[2,2]}$, and for identifiability reasons, we give preference to $\boldsymbol{\theta}^{[2,1]}$ by additionally shrinking $\boldsymbol{\theta}^{[2,2]}$ towards $\mathbf{0}$; this justifies the form of the block diagonal penalty matrix, $\boldsymbol{P}$, in (4) below (with the smoothing gap parameters $\lambda_{2,1}$ and $\lambda_{2,2}$, and the shrinkage parameter $\breve{\lambda}_{2,2}$). We now introduce the joint vectors of death counts and exposures: $\boldsymbol{d} = vec(\boldsymbol{d}^{[1]}, \boldsymbol{d}^{[2]})$ and $\boldsymbol{e} = vec(\boldsymbol{e}^{[1]}, \boldsymbol{e}^{[2]})$; the coefficient vector $\boldsymbol{\theta} = vec(\boldsymbol{\theta}^{[1]}, \boldsymbol{\theta}^{[2,1]}, \boldsymbol{\theta}^{[2,2]})$ is then estimated by the penalized GLM (or more correctly, the penalized quasi-log-likelihood) for $\boldsymbol{d}$ with regression matrix $\boldsymbol{B}$, offset $\log(\boldsymbol{e})$, log link, quasi-Poisson error and penalty matrix $\boldsymbol{P}$, where

$$\begin{aligned} \boldsymbol{B} &= \begin{bmatrix} \boldsymbol{B}_t \otimes \boldsymbol{B}_a & \mathbf{0} & \mathbf{0} \\ \boldsymbol{B}_t \otimes \boldsymbol{B}_a & \mathbf{1}_{n_t} \otimes \boldsymbol{B}_a & \boldsymbol{B}_t \otimes \mathbf{1}_{n_a} \end{bmatrix}, \\ \boldsymbol{P} &= blockdiag\left(\boldsymbol{P}^{[1]},\ \lambda_{2,1}\,\boldsymbol{\Delta}_a'\boldsymbol{\Delta}_a,\ \lambda_{2,2}\,\boldsymbol{\Delta}_t'\boldsymbol{\Delta}_t\ +\ \breve{\lambda}_{2,2}\,\boldsymbol{I}_{c_t}\right). \end{aligned} \tag{4}$$

The linear predictor (3) could be re-parameterized in the form

$$\begin{aligned} \log\left(\mathbb{E}\left[\boldsymbol{y}^{[2]}\right]\right) \quad = \quad & (\boldsymbol{B}_t \otimes \boldsymbol{B}_a)\,\boldsymbol{\theta}^{[1]} + (\mathbf{1}_{n_t} \otimes \mathbf{1}_{n_a})\,\theta^{[2]} \\ & + (\mathbf{1}_{n_t} \otimes \boldsymbol{B}_a)\,\boldsymbol{\theta}^{[2,a]} + (\boldsymbol{B}_t \otimes \mathbf{1}_{n_a})\,\boldsymbol{\theta}^{[2,t]}, \end{aligned} \tag{5}$$

where $(\mathbf{1}_{n_t} \otimes \mathbf{1}_{n_a})\,\theta^{[2]}$, $(\mathbf{1}_{n_t} \otimes \boldsymbol{B}_a)\,\boldsymbol{\theta}^{[2,a]}$ and $(\boldsymbol{B}_t \otimes \mathbf{1}_{n_a})\,\boldsymbol{\theta}^{[2,t]}$ represent respectively the constant component, the smooth age dependent component and the smooth year dependent component of the gap. Here $\boldsymbol{\theta}^{[1]}$ is smoothed as before, there is no constraint on $\theta^{[2]}$; $\boldsymbol{\theta}^{[2,a]}$ and $\boldsymbol{\theta}^{[2,t]}$ are smoothed and shrunk towards zero. These three components give an economical comparison between mortality tables in *similar* populations. With this representation, the model corresponding to each scenario of similarity (defined earlier in this section) is derived from (5) by keeping the appropriate components and taking away the other components.

3 Computational aspects and applications

The joint model for *similar* populations presented in section 2 is very computationally demanding if fitted with the standard GLM procedure, especially as the number of populations increases. In the general situation of p populations, we speed up the estimation as follow. (i) First observe that $\boldsymbol{B}$ is partitioned as $\boldsymbol{B} = [\mathbf{B_1} : \mathbf{B_2}]$, with $\mathbf{B_1} = \mathbf{1}_p \otimes \boldsymbol{B}_t \otimes \boldsymbol{B}_a$, and $\mathbf{B_2} = [\mathbf{0} : \boldsymbol{\Lambda}]'$, where $\boldsymbol{\Lambda}$ is a block diagonal matrix; a good use of this partition is efficient for solving the penalized iterative equations as well as for computing the diagonal elements of the hat matrix required for estimating the total effective dimension, the contribution of each population to the total effective dimension, and the dispersion parameters. (ii) Second, the Kronecker structure of each component in this partition together with the matrix structure of the data allows us to express the model as a Generalized Linear Array Model (GLAM), a high speed, low storage framework (Currie et al., 2006). Using (i) and (ii) simultaneously leads to very substantial gains in time. Finally, we choose the smoothing parameters by minimizing the scaled BIC, see Heuer (1997).

We now apply our approach to some mortality data taken from two sources: (a) The Human Mortality Database (HMD) and (b) the Continuous Mortality Investigation (CMI).
We start with the HMD data, and for illustration, we consider ages 30 to 90 and years 1960 to 2005. The residuals from our model applied to male and female mortality in Japan show that the model fits "well" (profile views for ages 70 and 75 are shown in Figure 1); hence we conclude that the dynamisms of mortality in these two populations are *similar*. By the same procedure, the plots and residuals indicate that the dynamisms of mortality for males in Japan and Netherlands are *different* (see profile views for ages 70 and 75 in Figure 2).
We now consider the data from the CMI. These data are of two types: data by lives and data by amounts. The first type consists of the number of claims (view as deaths by lives) and the number of policies at risk (viewed as exposure to risk by lives); the second type consists of the total amounts claimed (viewed as death by amounts) and the total amounts at risk (viewed as exposure to risk by amounts). These two types of data lead to the concept of mortality by lives and mortality by amounts. The joint model applied to these data shows that the dynamisms in the mortality by lives and by amounts are *similar in time* (profile views for ages 70 and 75 are shown in Figure 3). Moreover, our joint model appropriately captures the well known fact that mortality by lives is worse than that by amounts; our model corresponding to the *similar in time* scenario has a particular importance for forecasting in life insurance, since it ensures that the extrapolated trends in time for different ages for mortality by lives and by amounts do not cross each other.

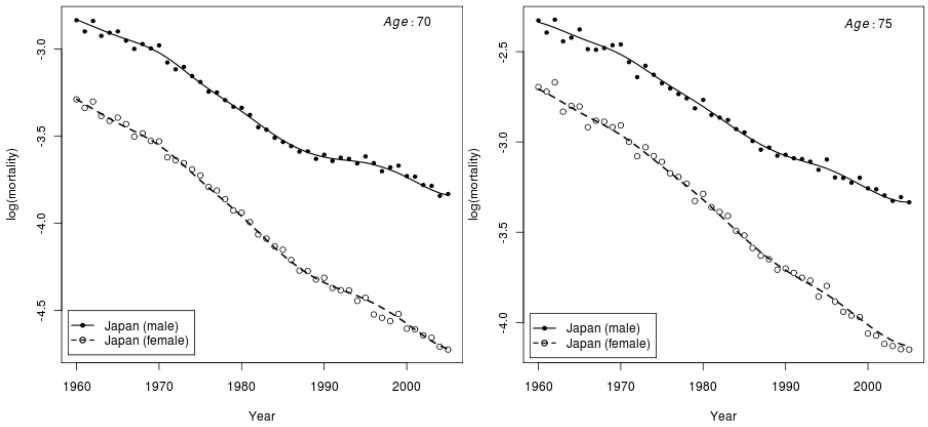


FIGURE 1. These profile views illustrate that the dynamisms in the male and female mortality in Japan are *similar*.

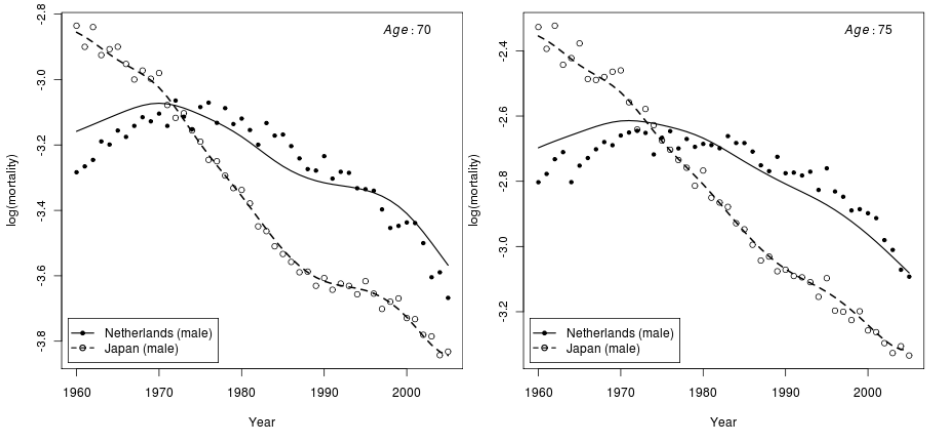


FIGURE 2. These profile views illustrate that the dynamisms in male mortality in Netherlands and in Japan are *different*.

4 Concluding remarks

In this paper we have proposed a class of joint models for classifying mortality tables. When two (or more) populations turn out to be similar in some way, our joint models lead to simple comparisons of these mortality tables. An additional attractive feature of our models is that, once the components are built, the fitting is reduced to the penalized scoring algorithm

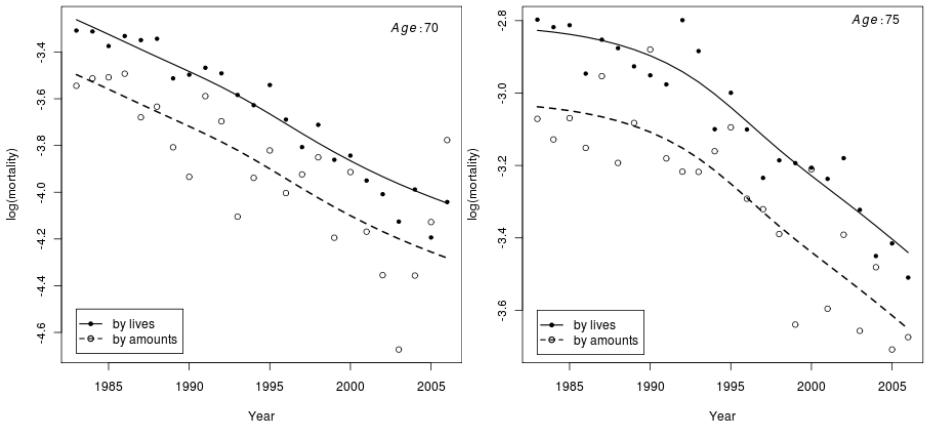


FIGURE 3. These profile views illustrate that the dynamisms in the CMI mortality by lives and by amounts are *similar in time*.

(with appropriate components). Furthermore, the order of the populations in our approach is not important; indeed taking population 2 (instead of population 1) as the reference leads to the same fit. We have approached the analysis of multiple mortality tables by fitting nested models. This has allowed us to compare such models by residual and graphical methods. Hypothesis testing is a more rigorous approach to such comparisons and our models give a platform for the development of these testing procedures. One problem that will need to be addressed is the very large power that our extensive datasets would give to any such test. This suggests that a Bayesian approach would be appropriate.

References

Currie, I.D., Durban, M., and Eilers, P.H.C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society, Series B*, **68**, 259-80.

Currie, I.D., Durban, M., and Eilers, P.H.C. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, **4**, 279-98.

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

Heuer, C. (1997). Modelling of time trends and interactions in vital rates using restricted regression splines. *Biometrics*, **53**, 161-177.

The identification and analysis of lip shape

Adrian Bowman[1], Denise Brown[1], Stanislav Katina[1]

[1] Department of Statistics, The University of Glasgow, Glasgow G12 8QQ

Abstract: Three-dimensional images of human faces can be captured by stereo-photogrammetry or other methods. From each of these images a number of facial anatomical landmarks are usually identified manually. This is a natural starting point for facial shape analysis but landmarks contain only a very small proportion of the data available from a captured image. Anatomically defined curves have the advantage of providing a much richer expression of facial shape. This is explored in the context of identifying the boundary between the upper lip and the surrounding skin, a feature which is important in a variety of surgical settings, including in particular the repair of a cleft lip and/or palate. Information on colour change and on surface shape can be used to identify these curves and the relative effectiveness of these two sources of information is investigated here.

Keywords: Curves; Landmarks; Shape analysis; Facial modelling.

1 Introduction

Statistical shape analysis has become a very important tool in medical imaging for the analysis of three-dimensional data on facial shape (Hood, 2003; Hammond, 2005). A natural starting point is to identify a set of points of correspondence, usually referred to as landmarks, across the population. Landmarks, placed around the face in an anatomically meaningful manner, are then analysed using well developed statistical methods (Dryden, 1998). However, the resolution of the underlying surface from stereo-photogrammetry systems, represented in mesh form, is very high and a landmark based analysis therefore takes into account only a very small proportion of the data available. Anatomical curves provide additional information about the shape of a face while retaining relatively low dimensionality. Methods for identification of such curves are therefore of considerable interest. Most of the approaches discussed in the literature are based on two-dimensional images while very different issues are raised in three dimensions.

A simple approach involves cutting the surface by a plane which contains two relevant landmarks. However, while this provides a helpful first approximation, there is no guarantee that the curve of interest lies in a plane and it would be more satisfactory to track the curve by its local properties. The upper lip boundary is an example of a curve which is of considerable anatomical interest. There is clearly potential for using both surface

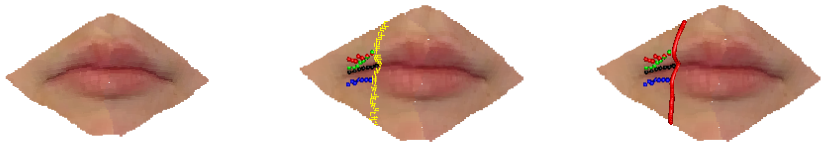


FIGURE 1. A mouth patch (left) with a strip of points identified by a cutting plane (middle) and the resulting principal curve (right) which tracks across the surface. The coloured points indicate lower, upper and midline positions identified by the methods discussed in the text.

shape and colour information to identify this curve. The left hand panel of Figure 1 shows the mouth region extracted from a face.

2 Methods

It is relatively straightforward to orient a mouth patch so that the x-axis runs across the lines of the mouth, the y-axis runs up the face and the z-axis represents depth and elevation of the facial surface. A principal components analysis of the points in the mouth patch is a convenient tool for achieving this. In order to reduce the complexity of the problem in identifying curves, the strategy adopted here is to scan the facial surface through strips running in the y-direction. This can be easily achieved by accepting into the strip any points whose value on the x-axis lies within some small tolerance of the x-value of interest. The middle panel of Figure 1 illustrates this by the yellow coloured points.

The concept of principal curves (Hastie & Stuetzle, 1989) offers a very convenient method for defining a curve which tracks along the facial surface. Principal curves are a non-linear and non-parametric version of principal components; the first principal curve is defined by the fact that the projection of the data onto this curve has largest possible variance. In the context of a facial surface, a principal curve based on a strip from a planar cut indexes a two-dimensional track, as illustrated by the curve in the right hand panel of Figure 1. The properties of this curve can then be used to identify the point at which the curve crosses the upper lip boundary.

At each observed three-dimensional location on the captured surface, information on colour is available as components on a red-green-blue colour scale. Graphical exploration shows that, surprisingly, it is green which carries the most information on the distinction between skin and lip tissue. Broadly speaking, the addition of green to red moves colour towards the yellower hue of skin from the redder hue of lip tissue. One way of identifying a point on the upper lip boundary is therefore to track along each

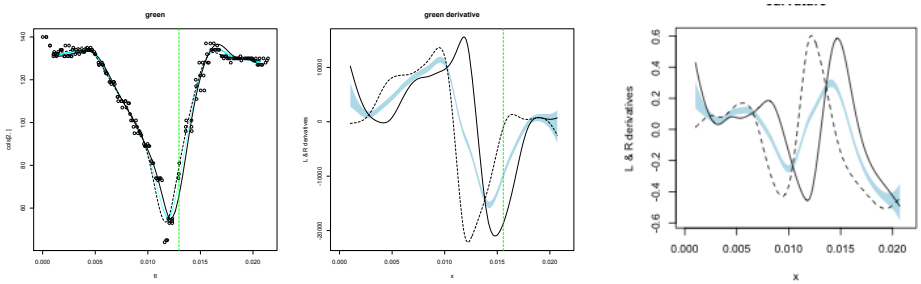


FIGURE 2. Curves to identify changes in green (left), green derivative (middle) and surface derivative (right) from a vertical strip across the mouth patch at a particular position. In each panel, the full and dashed lines represent right- and left-hand estimators respectively. The green vertical line shows the position of maximum standardised difference.

principal curve defined above and seek out change-points in the level of green. Bowman *et al.* (2006) describe how discontinuities in a nonparametric regression curve can be detected by comparing smooth estimators based on data to the left and data to the right. The left hand panel of Figure 2 illustrates this. However, although the level of green changes on moving from skin tissue to lip tissue, there is no discontinuity involved and the change continues as the principal curve moves into the shaded region where the lips meet. The pattern of the green level suggests that the lip boundary is characterised by a sharp change in derivative. The methods of Bowman *et al.* (2006) have therefore been modified to detect changes in slope rather than changes in level. The middle panel of Figure 2 shows that this is a very successful strategy, with the boundary point located in a position which is entirely consistent with visual judgement.

A similar approach can be adopted by using surface shape rather than colour. At each point on the principal curve, left hand and right hand smooth estimates based on the z-values for surface elevation can be used as the input to the slope change detection method. This is illustrated in the right hand panel of Figure 2 where the upper lip boundary point is again identified very effectively. Estimates of the lower lip boundary and the mid-line can also be constructed from the same information.

3 Application

Figure 3 shows the end result of this process on one illustrative case, with the estimate from green colour information on the left and from surface shape information on the right. The more variable nature of the estimate from colour is apparent and this general trend was evident in other cases where these methods were tried. A final step of constructing a boundary

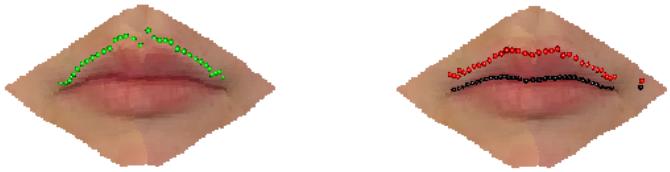


FIGURE 3. Lip curves identified by changes in green derivative (left) and surface derivative (right). The black line in the right hand panel identifies the mid-line.

through a further principal curve through the identified individual points is effective in reducing variability. However, the more effective nature of shape information is apparent in these initial points. It is likely that lighting effects from the photographs used in the image reconstruction will sometimes create difficulties for the slope change detection method.

Acknowledgments: The facial data were collected as part of a wider project carried out in collaboration with researchers from the Glasgow Dental Hospital and the Departments of Computing Science and Psychology at the University of Glasgow. The work of SK was supported by Wellcome Trust grant WT086901MA.

References

Bowman. A.W., Pope, A. & Ismail, B. (2006). Detecting discontinuities in nonparametric regression curves and surfaces. *Statistics & Computing* **16**, 377-390.

Dryden, I.L. and Mardi, K.V. (1998). *Statistical shape analysis*. Wiley: Chichester.

Hammond, P., Hutton, T. J., Allanson, J. E., Buxton, B., Campbell, L. E., Smith, J. C., Donnai, D., Smith, A. K., Metcalfe, K., Murphy, K. C., Patton, M., Pober, B., Prescott, K., Scambler, P., Shaw, A., Smith, A. C. M., Stevens, A. F., Temple, I. K., Hennekam, R. and Tassabehji, M. (2005). Discriminating power of localized three-dimensional facial morphology. *American Journal of Human Genetics* **77**, 999–1010.

Hastie, T. and Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association* **84**, 502–516.

Hood, C.A., Hosey, M.T., Bock, M.T., White, J., Ray, A. and Ayoub, A.F. (2005). Facial characterization of infants with cleft lip and palate using a three-dimensional capture technique. *Cleft Palate-Craniofacial Journal*, **41**, 27–35.

Signal estimation from observations with bounded random delays and packet dropouts

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

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

Abstract: In this paper, a recursive algorithm for the least-squares linear estimation problem is obtained in discrete-time linear stochastic systems with bounded random observation delays which can lead to bounded packet dropouts. The random delays are modelled by a sequence of Bernoulli variables with known distributions. The derivation of the proposed algorithm does not require full knowledge of the state-space model generating the signal to be estimated, but only the delay probabilities and the covariance functions of the processes involved.

Keywords: Least-squares estimation; Delayed observations; Packet dropouts.

1 Introduction

Originally, signal estimation problems were addressed under the assumption that the sensor data are transmitted over perfect communication channels. However, the use of communication networks for transmitting measured data motivates the need of considering eventual transmission delays and/or possible packet losses, due to numerous causes, such as random failures in the transmission mechanism or data inaccessibility at certain times. Assuming that the state-space model of the signal to be estimated is known, many results have been reported on systems with random delays and/or packet dropouts (see Sun (2009) and references therein).

On the other hand, when the state-space model of the signal to be estimated is not available, it is necessary to use alternative information, for example, about the covariance functions of the processes involved in the observation equation. In this context, the estimation problem from randomly delayed observations based on covariance information has been addressed in Caballero et al. (2009) and Linares et al. (2009), among others.

In this paper, the least-squares linear estimation problem in systems with bounded random measurement delays and packet dropouts is addressed. The proposed estimators only depend on the delay probabilities at each sampling time, but do not need to know if a particular measurement is delayed or updated. Moreover, the estimation algorithm is derived using

only covariance information. Consequently, considering the case of sensors with the same delay characteristics, the current study generalizes the results in Linares et al. (2009) to the case of measurements with bounded multiple-step random delays and packet dropouts.

2 Observation model and problem statement

Consider an n-dimensional signal process $\{z_k;\ k \geq 1\}$ with zero mean and autocovariance function $K_{k,s} = E[z_k z_s^T] = A_k B_s^T,\ s \leq k$, where A and B are known matrix functions. Assume that the measured output of the signal z_k, denoted by $\widetilde{y}_k$, is perturbed by a noise vector v_k; that is,

$$\widetilde{y}_k = z_k + v_k, \qquad k \geq 1$$

with $\{v_k;\ k \geq 1\}$ a zero-mean white sequence with known autocovariance function $E[v_k v_k^T] = R_k,\ \forall k \geq 1$.

The output $\widetilde{y}_k$ is transmitted at the sampling times $k, k+1, \cdots, k+D$ but, at each sampling time k, only one of the measurements $\widetilde{y}_{k-D}, \ldots, \widetilde{y}_k$ is processed; consequently, at time $k > D$, the processed measurement can be either delayed by $d = 1, \ldots, D$ sampling periods with a known probability $p_k^{(d)}$, or updated with probability $p_k^{(0)}$. At the initial time $k = 1$, the measured output of z_1 is always available ($p_1^{(0)} = 1$) and, hence, the processed measurement is equal to the real measurement, $y_1 = \widetilde{y}_1$. At any time $k \leq D$, the processed measurement only can be delayed by $d = 1, \ldots, k-1$ sampling periods, since only $\widetilde{y}_1, \ldots, \widetilde{y}_k$ are available. Also, it is assumed that the delays at different times are independent. Therefore, the following model for the processed measurements is considered

$$y_k = \sum_{d=0}^{\min(k-1,D)} \gamma_k^{(d)} \widetilde{y}_{k-d}, \qquad k \geq 1,$$

where, for $d = 0, 1, \ldots, D,\ \{\gamma_k^{(d)};\ k > d\}$ denote sequences of mutually independent Bernoulli variables with $P[\gamma_k^{(d)} = 1] = p_k^{(d)}$ and $\sum_{d=0}^{\min\{k-1,D\}} \gamma_k^{(d)} = 1$. Finally, we assume that, for each $d = 0, 1, \ldots, D$, the processes $\{z_k;\ k \geq 1\},\ \{v_k;\ k \geq 1\}$ and $\{\gamma_k^{(d)};\ k > d\}$ are mutually independent.

Our purpose is to find the least-squares (LS) linear estimator, $\widehat{z}_{k/L}$, of the signal z_k based on the observations $\{y_1, \ldots, y_L\}$. This estimator is the orthogonal projection of the vector z_k onto the n-dimensional linear space spanned by $\{y_1, \ldots, y_L\}$. We use an innovation approach, consisting of transforming the observation process $\{y_k;\ k \geq 1\}$ into an equivalent one (*innovation process*) of orthogonal vectors $\{\nu_k;\ k \geq 1\}$, which allows us to find the orthogonal projection by separately projecting onto each of the previous orthogonal vectors; thus, denoting $S_{k,i}^z = E[z_k \nu_i^T]$ and $\Pi_i = E[\nu_i \nu_i^T]$,

$$\widehat{z}_{k/L} = \sum_{i=1}^{L} S_{k,i}^{z} \Pi_i^{-1} \nu_i.$$

Hence, to obtain the signal estimators it is necessary to find previously an explicit formula for the innovations and their covariance matrices.

3 Innovation $\boldsymbol{\nu_k}$ and covariance matrix $\boldsymbol{\Pi_k}$

The innovation at time k is defined as $\nu_k = y_k - \widehat{y}_{k/k-1}$ where $\widehat{y}_{k/k-1}$ is the one-stage linear predictor of y_k, which clearly satisfies

$$\widehat{y}_{k/k-1} = \sum_{d=0}^{\min(k-1,D)} p_k^{(d)} \left(\widehat{z}_{k-d/k-1} + \widehat{v}_{k-d/k-1}\right), \quad k \geq 2; \qquad \widehat{y}_{1/0} = 0.$$

Since v_k is independent of $\{y_1, \ldots, y_{k-1}\}$, the one-stage predictor of the noise is $\widehat{v}_{k/k-1} = 0$. Consequently, the innovation is given by

$$\nu_k = y_k - p_k^{(0)} \widehat{z}_{k/k-1} - \sum_{d=1}^{\min(k-1,D)} p_k^{(d)} \left(\widehat{z}_{k-d/k-1} + \widehat{v}_{k-d/k-1}\right), \ k \geq 2; \ \nu_1 = y_1,$$

and it is necessary to obtain the signal predictor $\widehat{z}_{k/k-1}$ and the estimators $\widehat{z}_{k-d/k-1}$ and $\widehat{v}_{k-d/k-1}$ for $d = 1, \ldots, \min(k-1, D)$; that is, the filters and smoothers of the signal and noise, respectively.

Using the model assumptions, the covariance matrix Π_k is expressed as

$$\begin{aligned}
\Pi_k = & \sum_{d,d'=0}^{\min(k-1,D)} Cov(\gamma_k^{(d)}, \gamma_k^{(d')}) K_{k-d,k-d'} + \sum_{d,d'=0}^{\min(k-1,D)} p_k^{(d)} p_k^{(d')} P^{zz}_{k-d,k-d'/k-1} \\
& + \sum_{d=1}^{\min(k-1,D)} p_k^{(d)} \left(1 - p_k^{(d)}\right) R_{k-d} + p_k^{(0)} R_k + \sum_{d,d'=1}^{\min(k-1,D)} p_k^{(d)} p_k^{(d')} P^{vv}_{k-d,k-d'/k-1} \\
& + \sum_{d=0,d'=1}^{\min(k-1,D)} Cov(\gamma_k^{(d)}, \gamma_k^{(d')}) P^{zv}_{k-d,k-d'/k-1} + \sum_{d=1,d'=0}^{\min(k-1,D)} Cov(\gamma_k^{(d)}, \gamma_k^{(d')}) P^{vz}_{k-d,k-d'/k-1}
\end{aligned}$$

where $P^{zz}_{l,m/N} = E[(z_l - \widehat{z}_{l/N})(z_m - \widehat{z}_{m/N})]$, $P^{vv}_{l,m/N} = E[(v_l - \widehat{v}_{l/N})(v_m - \widehat{v}_{m/N})]$ and $P^{zv}_{l,m/N} = E[(z_l - \widehat{z}_{l/N})(v_m - \widehat{v}_{m/N})]$.

4 Recursive estimation algorithm

The filters and smoothers of the signal and noise are obtained by

$$\begin{aligned}
\widehat{z}_{k/L} &= \widehat{z}_{k/L-1} + S_{k,L}^{z} \Pi_L^{-1} \nu_L, \quad L \geq k, \quad k \geq 1, \\
\widehat{v}_{k/L} &= \widehat{v}_{k/L-1} + S_{k,L}^{v} \Pi_L^{-1} \nu_L, \quad L \geq k, \quad k \geq 1,
\end{aligned}$$

with initial conditions $\widehat{z}_{k/k-1} = A_k O_{k-1}$ and $\widehat{v}_{k/k-1} = 0$, respectively.

The vectors O_k satisfy the recursive relation

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

with

$$J_k = G_{B_k}^T - p_k^{(0)} r_{k-1} A_k^T - \sum_{d=1}^{\min(k-1,D)} p_k^{(d)} \left\{ r_{k-d} A_{k-d}^T + p_{k-d}^{(0)} J_{k-d} \Pi_{k-d}^{-1} R_{k-d} \right\}$$
$$- \sum_{d=2}^{\min(k-1,D)} p_k^{(d)} \sum_{i=1}^{d-1} J_{k-d+i} \Pi_{k-d+i}^{-1} \left\{ S_{k-d,k-d+i}^{zT} + S_{k-d,k-d+i}^{vT} \right\},$$

where $G_{B_k} = \sum_{d=0}^{\min(k-1,D)} p_k^{(d)} B_{k-d}$, and the matrix r_k is recursively obtained by

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

The gain matrices $S_{k,L}^z$ and $S_{k,L}^v$ are obtained by

$$S_{k,L}^z = \sum_{d=0}^{\min(L-1,D)} p_L^{(d)} P_{k,L-d/L-1}^{zz} + \sum_{d=1}^{\min(L-1,D)} p_L^{(d)} P_{k,L-d/L-1}^{zv}$$
$$S_{k,L}^v = \sum_{d=0}^{\min(L-1,D)} p_L^{(d)} P_{k,L-d/L-1}^{vz} + \sum_{d=1}^{\min(L-1,D)} p_L^{(d)} P_{k,L-d/L-1}^{vv}$$

where the covariance and cross-covariance matrices of the errors satisfies the following recursive expressions

$$P_{k,m/L}^{zz} = P_{k,m/L-1}^{zz} - S_{k,L}^z \Pi_L^{-1} S_{m,L}^{zT}, \quad L \geq k, m; \qquad k, m \geq 1,$$
$$P_{k,m/L}^{vv} = P_{k,m/L-1}^{vv} - S_{k,L}^v \Pi_L^{-1} S_{m,L}^{vT}, \quad L \geq k, m; \qquad k, m \geq 1,$$
$$P_{k,m/L}^{zv} = P_{k,m/L-1}^{zv} - S_{k,L}^z \Pi_L^{-1} S_{m,L}^{vT}, , \quad L \geq k, m; \qquad k, m \geq 1.$$

Acknowledgments: Research supported by grants MTM2008-05567 (Ministerio de Educación y Ciencia) and P07-FQM-02701 (Junta de Andalucía).

References

Caballero-Águila, R., Hermoso-Carazo, A., and Linares-Pérez, J. (2009). Least-squares polynomial estimation from observations featuring correlated random delays, *Methodology and Computing in Applied Probability*, doi: 10.1007/s11009-008-9117-z

Linares-Pérez, J., Hermoso-Carazo, A., Caballero-Águila, R., and Jiménez-López, J.D. (2009). Least-squares linear filtering using observations coming from multiple sensors with one or two-step random delay, *Signal Processing*, **89**, 2045–2052.

Sun, S. (2009). Linear minimum variance estimators for systems with bounded random measurement delays and packet dropouts, *Signal Processing*, **89**, 1457–1466.

Unscented filtering in nonlinear systems with uncertain observations and correlated noises

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

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

Abstract: An unscented filtering algorithm is proposed to estimate the state of a nonlinear discrete-time system from uncertain observations when the evolution of the state is governed by nonlinear functions of the state and noise, and the additive noise of the observation is correlated with that of the state. Accidental interruptions (uncertainty) in the observation process are modelled by binary random variables taking the values one and zero to indicate that the signal is present in the observation or that the observation is only noise, respectively.

Keywords: Uncertain observations; Unscented Kalman filter.

1 Introduction

Nonlinear filtering is an interesting research area in which many approaches have been developed, the most popular being the extended Kalman filter (Daum (2005)). This technique provides approximations of the mean and covariance of the signal which are accurate, at least, up to the first terms of their Taylor series expansions. Although the extended Kalman filter has been successfully applied to numerous nonlinear discrete-time systems, the use of truncated Taylor expansion yields some important drawbacks involving, on the one hand, the evaluation of the Jacobian matrices and, on the other, its instability. Among other nonlinear techniques, the unscented Kalman filtering (Julier and Uhlmann (2004)) is a relatively new one, which does not require the calculation of Jacobian matrices and improves the extended approach, providing approximations of the mean which are accurate up to the second term of its Taylor expansion. Several generalizations of the extended and the unscented Kalman filters have been proposed in Hermoso-Carazo and Linares-Pérez (2007) for nonlinear discrete-time systems with additive noises, using uncertain observations; from comparison of both techniques it is inferred that, also in the uncertainty case, the unscented filter is superior in performance.

The current paper concerns the estimation problem in nonlinear discrete-time systems from uncertain observations perturbed by additive noise cor-

related with the state noise. The unscented algorithm proposed extends to that in Hermoso-Carazo and Linares-Pérez (2007) in two directions. On the one hand, we consider a more general state model in which the noise is not necessarily additive and, on the other hand, the independence assumption between the state and observation noises is eliminated, thus addressing those situations in which the observation noise is correlated with the state.

2 Assumptions on the nonlinear model

Consider an n-dimensional discrete-time state process, $\{x_k;\ k \geq 0\}$, whose evolution is perturbed by a q-dimensional white noise, $\{w_k;\ k \geq 0\}$, and governed by known functions of the state and noise; that is:

$$x_{k+1} = f_k(x_k, w_k), \quad k \geq 0. \tag{1}$$

The observation model is specified by nonlinear functions of the state affected by the noises $\{v_k;\ k \geq 1\}$ and $\{\gamma_k;\ k \geq 1\}$:

$$y_k = \gamma_k h_k(x_k) + v_k, \quad k \geq 1 \tag{2}$$

where γ_k are binary random variables taking the value one if the observation y_k contains information on the state, and the value zero otherwise.
The first and second-order moments of the processes determining the evolution of the state and describing the observations are specified from the following hypotheses:
The initial state, x_0, is a random vector with mean $\overline{x}_0$ and covariance P_0.
The state and observation white noises are correlated zero-mean processes with covariances $E[w_k w_k^T] = Q_k, \quad E[v_k v_k^T] = R_k, \quad E[w_j v_k] = S_k \delta_{j,k-1}$.
The multiplicative noise $\{\gamma_k;\ k \geq 1\}$ describes the uncertainty in the observations and is a sequence of independent Bernoulli random variables with probabilities $P[\gamma_k = 1] = p_k$. The probability $1 - p_k$, named *false alarm probability*, represents the probability that the observed value at time k does not contain the signal.
Finally, x_0, $\Big(\{w_k;\ k \geq 0\}, \{v_k;\ k \geq 1\}\Big)$ and $\{\gamma_k;\ k \geq 1\}$ are mutually independent.

3 Estimation problem

Our purpose is to obtain an approximation of the least-squares optimal estimator, that is, the conditional mean $E\left[x_k/Y^k\right]$, with $Y^k = \{y_1, \ldots, y_k\}$. For this purpose, we use the unscented filtering procedure, which acts in two steps. First, taking into account the relationship (1), approximations of $E\left[x_k/Y^{k-1}\right]$ and $Cov\left[x_k/Y^{k-1}\right]$ are calculated from the conditional statistics of x_{k-1} and w_{k-1}; these approximations are then updated with

the new observation, y_k, to obtain the approximations of $E\left[x_k/Y^k\right]$ and $Cov\left[x_k/Y^k\right]$. The update is accomplished by the Kalman filter equations, which require the conditional statistics of y_k; hence, in view of (2), the correlation between x_k and v_k (or, more specifically, between w_{k-1} and v_k) must be taken into account in this step. These reasons led us to work jointly with the vectors x_{k-1}, w_{k-1} and v_k and hence, we define the augmented vectors $X_k = \left(\begin{array}{ccc} x_k^T, & w_k^T, & v_{k+1}^T \end{array} \right)^T$, $k \geq 1$. The problem is then reformulated as that of finding the filter of this augmented vector, $\widehat{X}_{k/k}$, whose first block-component provides the filter of the original state. The prediction and update steps are detailed in the following subsections.

3.1 Unscented algorithm: prediction step

For each $k \geq 1$ we start with approximations of the conditional mean, $\widehat{X}_{k-1/k-1}$, and the conditional covariance, $P^{XX}_{k-1,k-1/k-1}$, of X_{k-1} given Y^{k-1} The aim is to find approximations for the conditional mean and covariance of X_k given Y^{k-1} which, by the model hypotheses, are

$$\widehat{X}_{k/k-1} = \left(\begin{array}{c} \widehat{x}_{k/k-1} \\ 0 \\ 0 \end{array} \right), \quad P^{XX}_{k,k/k-1} = \left(\begin{array}{ccc} P^{xx}_{k,k/k-1} & 0 & 0 \\ 0 & Q_k & S_{k+1} \\ 0 & S_{k+1}^T & R_{k+1} \end{array} \right).$$

Hence, we only need the conditional statistics of $x_k = f_{k-1}(x_{k-1}, w_{k-1})$, which are approached from $\widehat{X}_{k-1/k-1}$ and $P^{XX}_{k-1,k-1/k-1}$ as follows:

- We consider a set of sigma-points $\left\{\chi_{i,k-1/k-1},\ i = 0, \ldots, 2N\right\}$ ($N = n + q + m$) and associated weights whose mean and covariance are exactly $\widehat{X}_{k-1/k-1}$ and $P^{XX}_{k-1,k-1/k-1}$ (see Julier and Uhlmann (2004) for details).
- Then, by defining $f^a_{k-1}(X_{k-1}) = f_{k-1}(x_{k-1}, w_{k-1}) = x_k$, the statistics of this function are approximated by those of the transformed sigma-points, $f^a_{k-1}(\chi_{i,k-1/k-1})$.

3.2 Unscented algorithm: update step

The approximations $\widehat{X}_{k/k-1}$ and $P^{XX}_{k,k/k-1}$ given in the previous section are now updated with the new observation, y_k, by using the Kalman filter equations. For this purpose, we need to approximate the mean and covariance of y_k given Y^{k-1}, as well as the conditional cross-covariance of y_k and X_k. Taking into account (2) and since $P\left[\gamma_k = 1/Y^{k-1}\right] = p_k$, these statistics are expressed in terms of those corresponding to $z_k = h_k(x_k)$ and v_k as

follows:

$$\begin{array}{l}\widehat{y}_{k/k-1} = p_k\widehat{z}_{k/k-1}\\ P^{yy}_{k,k/k-1} = p_k P^{zz}_{k,k/k-1} + p_k(1-p_k)\widehat{z}_{k/k-1}\widehat{z}^T_{k/k-1} + p_k P^{zv}_{k,k/k-1} + p_k P^{vz}_{k,k/k-1} + R_k\\ P^{Xy}_{k,k/k-1} = p_k P^{Xz}_{k,k/k-1} + P^{Xv}_{k,k/k-1}.\end{array} \tag{3}$$

Moreover, since z_k and v_k are conditionally independent of w_k and v_{k+1}, the cross-covariances $P^{Xz}_{k,k/k-1}$ and $P^{Xv}_{k,k/k-1}$ require only the cross-covariances of x_k with z_k and v_k; that is:

$$P^{Xy}_{k,k/k-1} = \begin{pmatrix} p_k P^{xz}_{k,k/k-1} + P^{xv}_{k,k/k-1} \\ 0 \\ 0 \end{pmatrix}.$$

The vectors z_k and x_k are both functions of x_k and, hence, their conditional statistics can be approximated from $\widehat{x}_{k/k-1}$ and $P^{xx}_{k,k/k-1}$ by considering the set of associated sigma-points, $\left\{\chi^x_{i,k/k-1},\ i=0,\ldots,2n\right\}$. However, since the vector v_k cannot be expressed in terms of X_k, its conditional statistics must be approximated from those of X_{k-1}. Thus by expressing $z_k = h_k(x_k)$ for $\widehat{z}_{k/k-1}$, $P^{zz}_{k,k/k-1}$ and $P^{zz}_{k,k/k-1}$ and $z_k = h_k\left(f^a_{k-1}(X_{k-1})\right)$ for $P^{zv}_{k,k/k-1}$, the required statistics are approximated by those of the sigma-points transformed by the above functions.

Finally, these statistics are substituted in (3) to obtain those of y_k, which are used in the following equations providing the filter and the filtering error covariance of X_k:

$$\begin{array}{l}\widehat{X}_{k/k} = \widehat{X}_{k/k-1} + P^{Xy}_{k,k/k-1}\left(P^{yy}_{k,k/k-1}\right)^{-1}\left(y_k - \widehat{y}_{k/k-1}\right)\\ P^{XX}_{k,k/k} = P^{XX}_{k,k/k-1} - P^{Xy}_{k,k/k-1}\left(P^{yy}_{k,k/k-1}\right)^{-1}P^{yX}_{k,k/k-1}.\end{array} \tag{4}$$

Acknowledgments: Research supported by grants MTM2008-05567 (Ministerio de Educación y Ciencia) and P07-FQM-02701 (Junta de Andalucía).

References

Daum, F. (2005). Nonlinear filters: beyond the Kalman filter, *IEEE Aerospace and Electronic Systems Magazine*, **20**(8), 57-69.

Julier, S.J. and Uhlmann, J. K. (2004). Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, **92**(3), 401-422.

Hermoso-Carazo, A. and Linares-Pérez, J. (2007). Different approaches for state filtering in nonlinear systems with uncertain observations. *Applied Mathematics and Computation*, **187**, 708–724.

Functional principal components models for high dimensional brain volumetrics

Vadim Zipunnikov[1], Brian Caffo[1], Ciprian Crainiceanu[1], Christos Davatzikos[2], Brian Schwartz[3]

[1] Department of Biostatistics, Johns Hopkins University
[2] Department of Radiology, University of Pennsylvania
[3] Department of Environmental Health, Johns Hopkins University

Abstract:

Keywords: Voxel-based morphometry; MRI; brain imaging data; functional principal components.

1 Introduction

In this manuscript we consider functional principal components analysis (FPCA) for very high dimensional data sets. We focus especially on settings where the data matrix can not be loaded into memory, let alone the high dimensional covariance matrices needed to perform FPCA. We discuss three main innovations. First, we make explicit the relationship between the data level singular value decomposition (SVD) and associated model-based estimates. Second, we demonstrate how calculations can be performed on low-memory systems in such a way that the data matrix never needs to be loaded into memory. Third, we apply the methods to a novel data set investigating cross-sectional variation in brain volumetrics. Here the data is thirteen million dimensional per subject, of which three million contain potentially important information. The exploratory nature of FPCA is particularly well suited to this analysis. Moreover, it is invariant to permutations of the high dimensional parameter space, which is important as the ordering of the three million observations is somewhat arbitrary. The analysis sheds considerable light on cross-sectional variation in brain shape.

The manuscript is laid out as follows. Section 1.1 outlines bran morphometry data while Section 1.2 describes our data set in specific. Section 2 covers relevant notation while Section 3 outlines the methodological connection between the SVD and model estimation. Section 4 covers computational issues while Section 5 discusses analysis of the data.

1.1 Brain morphometry data

Voxel-based morphometry (VBM) is a whole-brain method for studying localized changes is brain shape (see Ashburner and Friston 2000). Typically, VBM is performed by: coregistering brains to a common anatomical template, retaining transformation images in template space, analyzing the transformation images across subjects using voxel-wise regression models. In this work, we replace this final step with a more empirical analysis using functional principal components models.

1.2 Data

The data derive from an ongoing study of former organol-lead manufacturing workers (see Schwartz et al. 2007 and Caffo et al. 2007). In this data set, segmented gray and white matter T1 MRI images were registered to a standard template; we analyze the registered Jacobian of the transformation. The Jacobian images in question are referred to as "Regional Analysis of Volumes Examined in Normalized Space" (RAVENS) images (see Davatzikos et al. 2001). We smoothed the RAVENS images using a $10mm$ full width at half of the maximum Gaussian smoother. The data are a $256 \times 256 \times 198$ array per subject. However, efficient masking reduces the subject level data to roughly a 3 million length vector. We have 352 subjects, hence the data structure is $352 \times 3 \times 10^6$ in dimension.

2 Notation

Consider the analysis of a collection of functions, $\{Y_j(v)\}_{j=1}^J$, where Y_j is a real valued function and j indexes subject. We assume that the functions are observed on a regular grid, say $\{v_k\}_{k=1}^K$. Let Y be the observed data matrix, with (j,k) element $Y_j(v_k)$. We are concerned with instances where the number of observed points, K, is very large, on the order of several millions or larger. In our application the number of subjects is typically order of hundreds or thousands. Let $\mu(v)$ be $E[Y_j(v)]$ and μ be a K dimensional vector with element k $\mu(v_k)$. Let $\Sigma(v,v')$ be the covariance function $E[\{Y_j(v)-\mu(v)\}\{y_j(v')-\mu(v')\}]$ with estimator $\hat{\Sigma}(v,v') = \frac{1}{J}\sum_{j=1}^J \{y_j(v)-\bar{y}(v)\}\{y_j(v')-\bar{y}(v')\}$ and discretized estimator $\hat{\Sigma} = \frac{1}{J}Y'\{I-T(T'T)^{-1}T'\}Y$, where I is an identity matrix and T is a J dimensional vector of ones.
Consider the model

$$Y_j(v) = \mu(v) + \sum_{l=1}^{L} \lambda_l^{1/2}\psi_{jl}\phi_l(v), \tag{1}$$

where $\psi_l(v)$ and $\lambda_l^{1/2}$ are eigenfunctions and eigenvalues of $\Sigma(\cdot,\cdot)$, respectively. We further assume that ψ_{il} are uncorrelated random effects from a mean 0 variance 1 distribution.

3 Methods

In this section we will show that constructing the SVD of a detrended matrix Y is equivalent to performing the distribution-free FPCA. We start with estimating the mean image μ by the sample average $T(T'T)^{-1}T'Y$. Let $W = \{I - T(T'T)^{-1}T'\}Y$ be the detrended version of Y. Presume that the Spectral Value Decomposition of $W = UDV'$ where U is $J \times J$, D is a J dimensional diagonal matrix and V is a $K \times J$ matrix so that $U'U = V'V = I$. From the spectral decomposition of $W'W = VD^2V'$ we obtain the estimates of eigenvectors ϕ_l's as the columns of the matrix V. Variances λ_l are estimated by diagonal elements of matrix D^2. The number of included principal components, L, is chosen to make the explained variability $(\lambda_1 + \ldots + \lambda_L)/(\lambda_1 + \ldots + \lambda_J)$ large enough. Let $\psi_j = (\psi_{j1}, \ldots, \psi_{jL})'$ be the principal scores of the subject j. Then, the best linear unbiased predictors (BLUPs) of ψ_j is given by $\hat{\psi}_j = E(\psi_j|W) = D_L^{-1}(V_L'V_L)^{-1}V_L'W_j$ where V_L is a $K \times L$ matrix composed of the first L columns of matrix V and the diagonal matrix D_L contains first L elements of matrix D. The estimated BLUP, $\hat{\psi}_j$, boils down to a very simple and intuitive form. Indeed, the detrended image for subject j, W_j, can be represented through the SVD of W as $U_j'DV'$ where $U_j' = (U_{j1}, \ldots, U_{jJ})$ is the jth row of matrix U. Using this representation it can be easily shown that $\hat{\psi}_j = (U_{j1}, \ldots, U_{jL})$. In other words, the EBLUP of ψ_j is given by the first L components of the vector U_j' from the SVD of the detrended matrix W. Model 1 is a truncated Karhunen-Loeve (KL) expansion of image Y_j treated as a continuous stochastic process. Therefore, the results of this section provide novel geometric insights by linking together a standard tool of FDA, the KL expansion, and the cornerstone of linear algebra, spectral value decomposition.

Notice that a brute-force approach to estimating the eigenfunctions and principal component scores can not work in this application. The main reason is that it is infeasible to calculate the covariance operator $W'W = VD^2V'$, or even load the data matrix W into memory. Indeed, the size of the covariance operator is $K \times K$, which is nine trillion. Needless to say that trying to diagonalize such a matrix that cannot be loaded in the computer memory is a futile and wasteful enterprise. In the next section we develop a computationally efficient algorithm which can be implemented and runs in minutes on a laptop.

4 Computation

We have argued that a singular value decomposition of a detrended Y matrix yields estimates of all of the Model 1. However, calculating the SVD is difficult in settings where Y itself cannot be loaded into memory. Let $Y = [Y^{(1)}, \ldots, Y^{(R)}]$ where each $Y^{(r)}$ is a submatrix of Y. Then the

detrended matrix can be partitioned in a similar way

$$\begin{aligned} W &= [\{I - T(T'T)^{-1}T'\}Y^{(1)}, \ldots, \{I - T(T'T)^{-1}T'\}Y^{(R)}] \\ &\equiv [W^{(1)}, \ldots, W^{(R)}]. \end{aligned}$$

Notice these calculations can be done serially, by loading each $Y^{(r)}$ one at a time. Further notice that from SVD $W = UDV'$ the covariance matrix $J\hat{\Sigma} = W'W = VD^2V'$ and hence the columns of V contain the eigenvectors, $\psi_k(v)$, for Model 1. Moreover, element (i, k) of U contains the best linear prediction estimates for ψ_{ik}, while diagonal elements of D contain $\lambda_k^{1/2}$.

However, since Σ contains K^2 elements, it is far too large to be loaded into memory, an alternative method for computation is needed. Consider that $WW' = UD^2U' = \sum_{r=1}^{R} W^{(r)}(W^{(r)})'$ is on the manageable order of J^2. Notice also that WW' can be calculated iteratively by summing over the $W^{(r)}$ so that W does not have to be loaded into memory. Hence we presume that U and D can be calculated via an eigenvalue decomposition of WW'. Then U can be calculated via serial computations as $V = D^{-1}U'W = [D^{-1}U'W^{(1)} \ldots D^{-1}U'W^{(R)}]$.

To summarize we have outlined a procedure that computes the relevant estimates of Model 1 via an SVD. The SVD is performed in a manner that only requires a number of operations that is linear in the larger dimension of the data matrix. Furthermore, we showed how the SVD can be calculated without loading the entire data matrix into memory. Hence, the analysis scales to nearly arbitrary large parameter spaces on very modest computing infrastructures.

5 Results

We apply cross-sectional functional data analysis methods to the Lead Study to RAVENS images. We consider cross-sectional eigen-analysis of the white matter and grey matter RAVENS images, each having been smoothed with a 10mm FWHM Gaussian smoother. To implement the approach we created 100 matrices consisting of subjects (rows) and voxels (columns) where the roughly 30 thousand voxels in each matrix represent a parcelation of the roughly 3 million non-background voxels in template space. The entire analysis performed in Matlab 2010a took about 9 minutes on a PC with quad core 2.6Gz processor and 6 GB of RAM. Figure 1 shows the amount of the explained variability explained by the first L components, $(\lambda_1 + \ldots + \lambda_L)/(\lambda_1 + \ldots + \lambda_J)$. Figure 2 contains the average image μ as well as the first three components ϕ_1, ϕ_2, and ϕ_3.

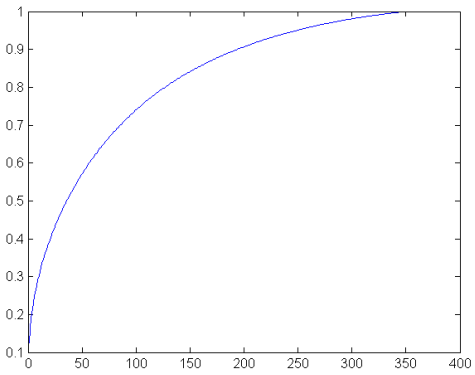


FIGURE 1. The percentage of the explained variability $(\lambda_1 + \ldots + \lambda_L)/(\lambda_1 + \ldots + \lambda_J)$ against L.

References

Ashburner, J. and Friston, K. (2000). Voxel-based morphometry, the methods. *Neuroimage*, **11**, 805-821.

Caffo, B., Chen, S., Stewart, W., Bolla, K. Youssem, D., Davatzikos, C., Schwartz, B. (2007). Are brain volumes based on magnetic resonance imaging mediators of the associations of cumulative lead dose with cognitive function?. *American Journal of Epidemiology*, **167**, 429-437

Di, C., Crainiceanu, C., Caffo, B. and Punjabi, N. (2009). Multilevel functional principal component analysis. *Annals of Applied Statistics*, **3**, 458-488.

Davatzikos, C., Genc, A., Xu, D. and Resnick, S. (2001). Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy. *Neuroimage*, **14**, 1361-1369.

Schwartz, B., Chen, S., Caffo, B., Stewart, W., Bolla, K., Youssem, D., Davatzikos, C. (2007). Relations of brain volumes with cognitive function in males 45 years and older with past lead exposure. *Neuroimage*, **37**, 633-641.

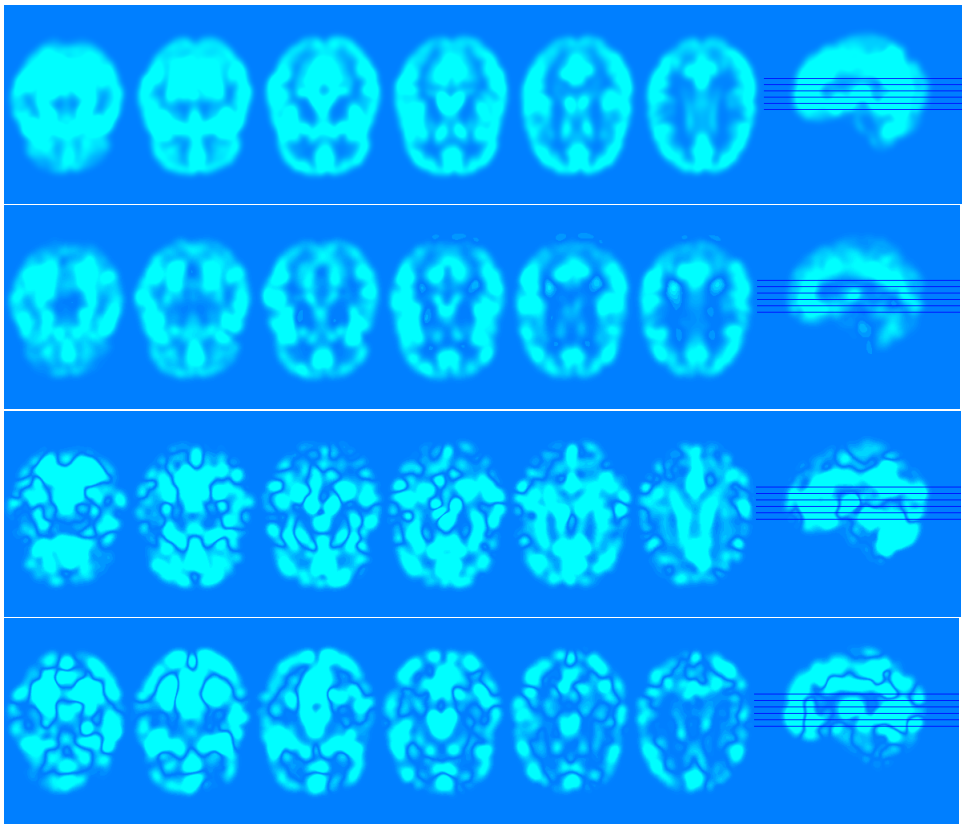


FIGURE 2. Average image μ and the first three components ϕ_1, ϕ_2, and ϕ_3.

Sums of Smooth Exponentials

Carlo G. Camarda[1], Paul H. C. Eilers[2], Jutta Gampe[1]

[1] Max Planck Institute for Demographic Research, Rostock, Germany.
camarda@demogr.mpg.de, gampe@demogr.mpg.de
[2] Department of Biostatistics, Erasmus Medical Centre, Rotterdam,
The Netherlands. p.eilers@erasmusmc.nl

Abstract: Standard smoothing procedures are challenged if the smoothness of the curve to be fitted varies strongly over the domain. A model is presented that conceptualizes the structure of such data as a sum of smooth components, which are modelled on the log scale and are then additively combined. The resulting estimation algorithm leads to a penalized composite link model. The approach is illustrated by three data examples.

Keywords: Additive Components; Composite Link Model; Logarithmic scale.

1 Introduction

Some functions look more smooth on a logarithmic scale than on the linear scale. A typical example is a Gaussian peak, evaluated over a larger domain. To the right and to the left of the peak the function is flat and very smooth, but the middle of the peak shows rapid changes. If we measure roughness by squared second or third order derivatives, large values occur near the middle of the peak and small values in the tails. In contrast, the third derivative of the logarithm (which is a parabola) vanishes everywhere, implying complete smoothness. In a sum of several Gaussian peaks at separated positions, with different widths and heights, very smooth, namely essentially flat regions and the highly non-smooth peaks alternate. When such data are smoothed by simple penalty methods, typically no good compromise between rounding the peaks and spurious wiggles in the flatter parts is possible. One solution is to introduce a locally adaptive penalty, or equivalently, use variable weights for the data. A number of papers on this subject have appeared.
Here we propose an alternative approach, in which we model data as a sum of exponentials of (very) smooth parametric or semi-parametric components. The estimation algorithms leads us to special cases of the composite link model.
We illustrate our approach with three data sets. One gives the logarithm of mortality of males in Italy in 1975. Mortality generally increases quite smoothly after about age 10, but it shows a steep decline between birth and

this minimum, due to infant and child mortality. Also for ages around 20 a so called accident-hump is superimposed on the general mortality increase. We model the data as a sum of three components: A smooth trend (for all ages), a parametric component for infant mortality, which is the exponential of a linear (downward sloping) function of age, and the exponential of a smooth component to capture the accident hump.
The second dataset is a two-dimensional generalization of the first example. We analyze how age pattern changes over time for the men in England and Wales between 1970 and 1990. Both datasets are taken from the Human Mortality Database (2010).
The third data sets contains a part of an X-ray diffraction spectrum (Davies et al, 2008). It shows a slowly varying baseline plus some isolated peaks. Of the latter the logarithms are modeled.
Note that in all the examples we effectively split our data into several components, which have a clear meaning: we quantify the amount and the speed of decay of infant mortality, as well as the contribution of the accident hump; and we can accurately determine heights, widths and positions of the diffraction peaks, corrected for the drifting baseline.

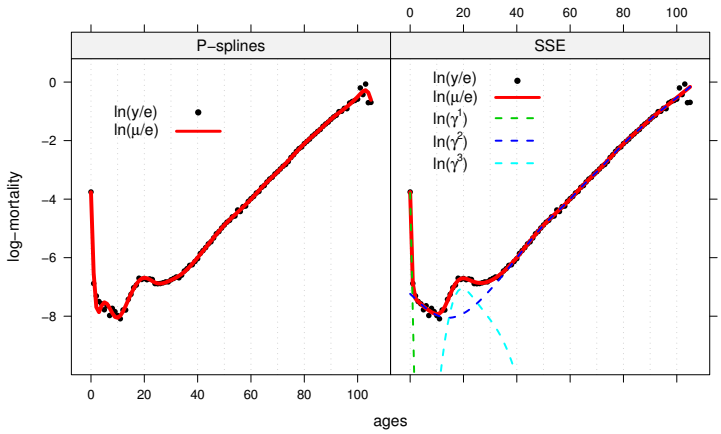


FIGURE 1. Actual and fitted death rates by P-splines (Eilers and Marx, 1996) and SSE model (cf. Section 2), log-scale. Italy, males, 1975, ages 0-105.

2 Sums of Smooth Exponentials

We formulate the model for a one-dimensional Poisson regression problem, as needed for the first mortality example. Let $\boldsymbol{e}$ be the m-dimensional vector

of exposures at the m ages considered, and $\boldsymbol{y}$ be the corresponding vector of numbers of deaths. The actually observed death counts are assumed to be realizations from a Poisson distribution, $\boldsymbol{y} \sim \mathcal{P}(\boldsymbol{\mu})$. The expected values $\boldsymbol{\mu}$ are the product of exposures $\boldsymbol{e}$ and actual death rates at the respective ages.
We model $\boldsymbol{\mu}$ as a composition of K components, each of length m, such that $\boldsymbol{\mu}$ can be written as $\boldsymbol{\mu} = \boldsymbol{C}\boldsymbol{\gamma}$, with $\boldsymbol{\gamma}' = [\boldsymbol{\gamma}^1 : \cdots : \boldsymbol{\gamma}^k : \cdots : \boldsymbol{\gamma}^K]$ and each $\boldsymbol{\gamma}^k \in \mathbb{R}^m$. The matrix $\boldsymbol{C}$ additively combines the $\boldsymbol{\gamma}^k$ and also incorporates the exposures:

$$\boldsymbol{C} = \mathbf{1}_{1,K} \otimes \mathtt{diag}(\boldsymbol{e})\,. \tag{1}$$

For each component we assume that the discrete sequence $\boldsymbol{\gamma}^k$ can be written as $\boldsymbol{\gamma}^k = \exp(\boldsymbol{X}^k\boldsymbol{\beta}^k)$, $k = 1, \ldots, K$. The design matrices $\boldsymbol{X}^k$ can represent parametric or non-parametric structures. In this way the composed mean $\boldsymbol{\mu}$ can be viewed as a sum of K exponential components, which eventually can be smooth. Hence we call this a Sum of Smooth Exponentials (SSE) model.

3 A Composite Link Model Approach

The SSE model can be embedded in a composite link model framework (Thompson and Baker, 1981). We therefore can estimate all components simultaneously using an adjusted iterative re-weighted least squares (IRWLS) algorithm. Given (1), the modified design matrix is given by

$$\breve{\boldsymbol{X}} = \left(\breve{\boldsymbol{X}}^1 : \cdots : \breve{\boldsymbol{X}}^k : \cdots : \breve{\boldsymbol{X}}^K\right) \tag{2}$$

where, for each component, we have $\breve{x}^k_{ij} = e_i\, x^k_{ij}\, \frac{\gamma^k_i}{\mu_i}$.
The structures of (1) and (2) lead to a major computational advantage: Both $\boldsymbol{\gamma}$ and $\breve{\boldsymbol{X}}$ can be constructed independently for each k. The IRWLS can be thus written as

$$(\breve{\boldsymbol{X}}'\tilde{\boldsymbol{W}}\breve{\boldsymbol{X}} + \boldsymbol{P})\tilde{\boldsymbol{\beta}} = \breve{\boldsymbol{X}}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}}\,, \tag{3}$$

where $\tilde{\boldsymbol{W}} = \mathtt{diag}(\tilde{\boldsymbol{\mu}})$ and $\tilde{\boldsymbol{z}} = \tilde{\boldsymbol{W}}^{-1}(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}) + \breve{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}$. The vector $\tilde{\boldsymbol{\beta}}$ concatenates the coefficients $\tilde{\boldsymbol{\beta}}^k$ of the K components. The penalty matrix $\boldsymbol{P}$ is a block diagonal matrix:

$$\boldsymbol{P} = \mathtt{diag}(\boldsymbol{P}^1, \ldots, \boldsymbol{P}^k, \ldots, \boldsymbol{P}^K),$$

where, in the case of a smooth component, $\boldsymbol{P}^k = \lambda^k \boldsymbol{D}_d^{k'} \boldsymbol{D}_d^k$ and $\boldsymbol{D}_d^k$ is the matrix that computes d-th order differences for the coefficients $\boldsymbol{\beta}^k$. The smoothing parameter λ^k controls the roughness of the vector $\boldsymbol{\gamma}^k$. Alternatively, if a parametric model for the component is chosen, $\boldsymbol{P}^k$ is a matrix of zeros. To choose the optimal amount of smoothness we minimize the Bayesian Information Criterion where the effective dimension is the trace of the hat-matrix from the estimated system in (3).

3.1 SSE in Two Dimensions

The SSE model can be generalized to a two-dimensional setting. For mortality data let $\boldsymbol{Y} = (y_{ij})$ be the matrix of deaths at age $i, i = 1, \ldots, m$ and year $j, j = 1, \ldots, n$. We arrange the matrix of deaths by column order into a vector $\boldsymbol{y}$. Likewise we arrange the matrix of exposures $\boldsymbol{e} = \texttt{vec}(\boldsymbol{E})$. We can thus directly employ (3) with modified versions of design and penalty matrices.

If a smooth two-dimensional surface is chosen for component k, the corresponding design matrix is: $\boldsymbol{X}^k = \boldsymbol{B}_y^k \otimes \boldsymbol{B}_x^k$. $\boldsymbol{B}_y^k \in \mathbb{R}^{n \times r_y}$ and $\boldsymbol{B}_x^k \in \mathbb{R}^{m \times r_x}$ denote two univariate B-splines over the two domains. The associated penalty term is given by

$$\boldsymbol{P}^k = \lambda_x^k \, \boldsymbol{I}_{r_y} \otimes \boldsymbol{D}_x^{k'} \boldsymbol{D}_x^k \; + \; \lambda_y^k \, \boldsymbol{D}_y^{k'} \boldsymbol{D}_y^k \otimes \boldsymbol{I}_{r_x} \, , \tag{4}$$

where λ_x^k and λ_y^k are the smoothing parameters for the two dimensions of component k (Currie et al., 2004).

Alternatively, a generalized additive model can be incorporated for a given component. In this case the design matrix is $\boldsymbol{X}^k = [\boldsymbol{1} : \boldsymbol{B}_x^k : \boldsymbol{B}_y^k]$ with a block-diagonal matrix for the penalty term: $\boldsymbol{P}^k = \texttt{diag}(0, \boldsymbol{P}_x^k, \boldsymbol{P}_y^k)$. Both $\boldsymbol{P}_x^k$ and $\boldsymbol{P}_y^k$ are built as in the one-dimensional case. Since each of the columns of $\boldsymbol{B}_x^k$ and $\boldsymbol{B}_y^k$ sum to one, a ridge penalty is needed for a GAM component. A penalty of 10^{-4} worked well in our examples (Durban et al., 2002).

4 Applications

4.1 Mortality Data in 1D

Figure 1, left panel shows a straightforward P-spline smooth for the mortality data given in the introduction. The right panel shows the results obtained from the SSE model. We use a combination of a single exponential function for the infant mortality, a penalized B-splines basis for the overall mortality, and a third component for the accident hump, whose support is restricted to ages 10 to 40 and is modelled by P-splines as well.

4.2 Mortality Data in 2D

Mortality development of English males from 1950 to 1990 and ages 0 to 105 is used for our second example. Figure 2 presents the actual and fitted values over age for selected years. For the infant mortality component we employed a series of exponential functions over years. A two-dimensional penalized surface is used for describing the overall mortality pattern. The development of the accident hump over the period is modelled by a generalized additive model, restricted to ages between 3 and 50. In this example smoothing parameters were subjectively chosen.

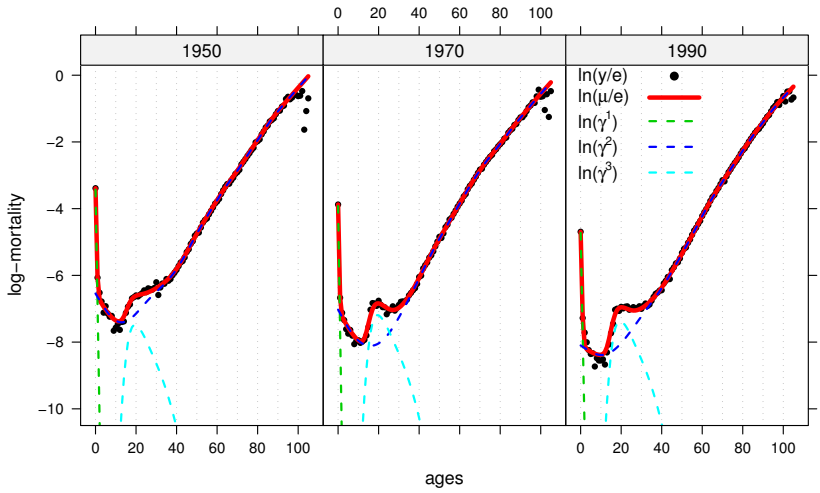


FIGURE 2. Actual and fitted death rates by two-dimensional SSE model, log-scale for selected years. England and Wales, males, 1950-1990, ages 0-105.

4.3 Indium Oxide Application

Figure 3 presents another application of the SSE model. Photon counts from Indium oxide doped with tin are plotted against a series of angles of diffraction. Clearly the dataset asks for a baseline trend and two peaks. The SSE model can disentangle these three components, which were all described by penalized B-splines. Moreover, the SSE model allows to efficiently model each peak in its neighborhood only.

5 Concluding Remarks

In this paper we present a model for smoothing data which vary strongly over their domain. Instead of searching for a global smoother, the proposed SSE model decomposes the overall structure in a sum of smooth components. The estimation procedure employs a P-spline approach within a composite link model. Parametric and non-parametric models can be used for describing each component; and both one- and two-dimensional settings have been introduced. This allows the SSE model to properly portray diverse datasets as shown in the presented applications.

References

Currie I.D., Durban M. and Eilers P.H.C. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*; **4**, 279-298

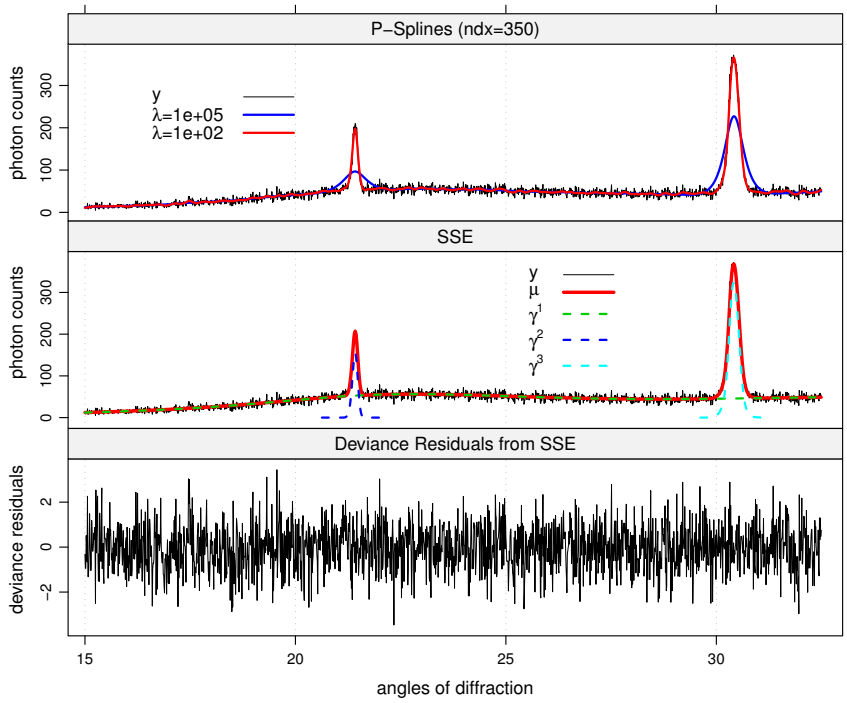


FIGURE 3. Actual and fitted counts by P-splines and SSE model (top panels). Deviance residuals from SSE model (bottom panel). Indium oxide data.

Davies, P.L., Gather, U., Meise, M., Mergel, D. and Mildenberger, T. (2008). Residual based localization and quantification of peaks in X-ray diffractograms. *Annals of Applied Statistics*. **2**, 861-886.

Durban M., Currie I.D., Eilers P.H.C. (2002). Using P-splines to smooth two-dimensional Poisson data. *Proc. 17th IWSM*, Chania, Greece, 207-214.

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

Human Mortality Database (2010). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at `www.mortality.org`.

Thompson, R. and Baker, R.J. (1981). Composite Link Functions in Generalized Linear Models. *Applied Statistics*, **30**, 125-131.

Modelling censored data with the skew-normal distribution

Rosa Capobianco[1], Jane Hutton [2], Elena Stanghellini[3]

[1] Department of Cultural and Educational Studies, University of Roma Tre, Roma, I-00185 Italy
[2] Department of Statistics, University of Warwick, Coventry, CV5 6AB, UK
[3] D.E.F.S., University of Perugia, Via A. Pascoli, Perugia, I-06100 Italy, email: elena.stanghellini@stat.unipg.it

Abstract: Motivated by a real data example, coming from a clinical trial to compare four treatments on severely sprained ankles, we extend the normal model for censored data, also known as Tobit model, to accomodate asymmetry in the distribution of the residuals. The proposed model makes use of the skew-normal distribution, a distribution that includes the normal one. The theoretical features of the model are investigated. These involve the analysis of the behaviour of the score functions and the formulation of a R-squared type of measure to evaluate the capacity of the model to represent the data.

Keywords: censored data, skew distributions, skew-normal distribution

1 Introduction

Regression models with censored responses arise in many applications, ranging from econometrics to biometrics. In the medical context, censored data can occur when the aim is to evaluate the success of a treatment, the outcome of which cannot be directly observed. In this situation, measurements are often derived through questionnaires. Each item of the questionnaire aims to asses one aspect of health or ability, by asking the patient to give a score. Typically, the scores have a finite range. The overall quality of life score is formed as a weighted sum of different scores. For repeated measurements, later scores tend to cluster at one end of the range, usually associated to the positive outcome, giving rise to a censoring mechanism at the threshold for adequate recovery.
A standard way to deal with censored responses is through the Gaussian censored model, also known as Tobit model (Tobin, 1956). For right censored data the model is:

$$Y^* = \beta^T x + \eta \tag{1}$$

with η distributed as $N(0, \sigma)$ and β a vector of unknown regression coeffi-

cients. The observed variable Y is defined as:

$$\begin{cases} Y = Y^* & \text{if } Y^* < T \\ Y = T & \text{otherwise} \end{cases} \tag{2}$$

where T is a known threshold. Note that this model may arise also in a sample-selection context, which is not presented here.

To distinguish between the two situations, it is common in the literature to address the above model as censored Tobit for corner solution outcomes (Wooldridge, 2003, ch. 16). The assumptions underlying the model, i.e. homoscedasticity and normality of the residuals, are often too strong to be satisfied and recent contributions in the literature have focussed on estimating the model when some of these assumptions are violated. In particular, since the Gaussian distribution appears to be unsuitable in many applied contexts, distribution-free methods have been specified, see Powell (1994) for a summary.

One problem when dealing with repeated measurements of quality of life scores is that, as the mean increases or decreases over time, skewness increases as well. This effect may be due to the presence of a boundary. Therefore, the underlying assumption that the residuals follow a Gaussian distribution should be abandoned in favour of a distribution that allows some degree of skewness in the data.

Driven by a real-data problem, related to a randomized controlled trial of four treatments for acutely sprained ankle, in this paper we present a model that includes the Tobit model as a special case. The model makes use of the skew-normal distribution (see Azzalini, 2005). More precisely, we will assume that η is distributed as $SN(0, \sigma, \alpha)$, with α the skewness parameter. The interest on the SN model lies in the fact that for $\alpha = 0$ the SN becomes a normal $N(0, \sigma)$ distribution. In Hutton and Stanghellini (2009), among other analyses, the censored skew-normal model has been implemented and maximum likelihood estimators, based on numerical optimization of the log-likelihood, have been derived. In this paper, we investigate the theoretical features of the model.

In section 2, the skew-normal censored model is presented. Testing whether $\alpha = 0$ is a crucial issue. A peculiarity of the SN model is that for $\alpha = 0$ the score functions become proportional and therefore the information matrix is not full rank. In section 3 the behaviour of the score functions of the proposed model when $\alpha = 0$ is investigated. Moreover, the construction of a proper R-squared index of the accuracy of the model is also outlined. In section 4 we present some final remarks.

2 The skew-normal censored model

Let Y be a random variable with density function:

$$f(y) = \frac{2}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\Phi(\alpha\sigma^{-1}(y-\mu)).$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denotes the standard normal density and distribution function. The distribution is known as skew-normal, in short $SN(\mu, \sigma, \alpha)$. It was initially studied by O'Hagan and Leonard (1976) and Azzalini (1985). For $\alpha = 0$, the distribution takes the form of normal distribution.
Under the assumption that η in (1) follows a $SN(0, \sigma, \alpha)$, the likelihood for a generic i-th observation then becomes:

$$L_i = \left[\frac{2}{\sigma}\phi\left(\frac{y_i - \mu_i}{\sigma}\right)\Phi(\alpha\sigma^{-1}(y_i - \mu_i))\right]^{1-a_i} \times [2\Phi_2(0, \sigma^{-1}(T - \mu_i); -\rho)]^{a_i}$$

in which $a_i = I[y_i^* \geq T]$, $\mu_i = \beta^T x_i$ and $\Phi_2(\cdot)$ denotes the distribution function of a standard bivariate normal with $-\rho$ as a correlation coefficient, $\rho = \alpha/\sqrt{1+\alpha^2}$. The expression of the second element comes from the cdf of a skew-normal distribution (see Azzalini, 1985, for details).

3 The score functions

We here derive the score function of the i-th observation drawn from model (1) w.r. to μ_i, σ, α. These are:

$$S(\mu_i) = \frac{\partial \log L_i}{\partial \mu_i} = (1 - a_i)\left\{\frac{z_i}{\sigma} - \frac{\alpha}{\sigma}w(\alpha z_i)\right\} - \frac{a_i\phi(u_i)\Phi(\alpha u_i)}{\sigma\Phi_2(0, u_i)}$$

where $z_i = \frac{y_i - \mu_i}{\sigma}$, $u_i = \frac{T - \mu_i}{\sigma}$ and $w(\alpha z_i) = \frac{\phi(\alpha z_i)}{\Phi(\alpha z_i)}$. Furthermore,

$$S(\sigma) = \frac{\partial \log L_i}{\partial \sigma} = (1 - a_i)\left\{-\frac{1}{\sigma} + \frac{z_i^2}{\sigma} - \alpha\frac{z_i}{\sigma}w(\alpha z_i)\right\} - \frac{a_i u_i\phi(u_i)\Phi(\alpha u_i)}{\sigma\Phi_2(0, u_i)}$$

and

$$S(\alpha) = \frac{\partial \log L_i}{\partial \alpha} = (1 - a_i)z_i w(\alpha z_i) - \frac{a_i\phi(\sqrt{1+\alpha^2}u_i)}{(1+\alpha^2)\sqrt{2\pi}\Phi_2(0, u_i)}.$$

It is easy to verify that for $\alpha = 0$, $S(\mu_i)$ and $S(\alpha)$ reduce to the score function of the Tobit model. However, in this situation, $S(\alpha)$ is proportional to $S(\mu_i)$, i.e. $S(\alpha)_{|\alpha=0} = \sigma\sqrt{\frac{2}{\pi}}S(\mu_i)_{|\alpha=0}$. This implies that under the null hypothesis $H_0 : \alpha = 0$ the information matrix becomes singular and the asymptotic distribution of the LRT is non-standard. A possible solution to overcome this problem could be the reparametrization of the model to apply the results in Rotnizky et al. (2000). The role of the centered parametrization (Azzalini, 2005) will also be investigated.
Furthermore, let $\hat{y}_i$ be the estimate of $E[Y \mid x = x_i]$. Under the assumptions of (1) and (2), this is:

$$P(Y^* < T \mid x = x_i)E[Y^* \mid x = x_i, Y^* < T] + TP(Y^* \geq T \mid x = x_i).$$

A measure of the accuracy of the model can be also based on the relationship between $\hat{y}_i$ and y_i, as in the R-squared index.

4 Final remarks

The assumption that the residuals of the Tobit model follow a Gaussian distribution has been questioned in a number of studies. The derivations in this paper specify a larger model that includes the Tobit model, and allows for asymmetry in the data. Some aspects of inference using the model need to be investigated. Those involve the behaviour of the information matrix and the construction of an appropriate R-squared measure.

Acknowledgments: Special Thanks to Professor Sallie Lamb and the CAST team for encouraging these investigations.

References

Azzalini, A (1985). A class of distributions which includes the Normal Ones. *Scandinavian Journal of Statistics* **12**, 171–178.

Azzalini, A. (2005). The Skew-normal Distribution and Related Multivariate Families. *Scandinavian Journal of Statistics* **32**, 159–188.

Hutton, J.L., and Stanghellini, E. (2009). Modelling health scores with the skew-normal distribution. CRiSM Research Report, n. 39.

O'Hagan, A., and Leonard, T. (1976). Bayesian Estimation subject to uncertainty about parameter constraints. *Biometrika*, **63**, 201-202.

Powell, J.L. (1994). Estimation of semi-parametric models. In *Handbook of Econometrics*, 4. Eds: R.F. Engle and D. McFadden. Amsterdam: North Holland, 2443-2521.

Rotnitzky A., Cox, D.R., Bottai, M., and Robins, J. (2000). Likelihood-based inference with singular information matrix. *Bernoulli* 6, 395–401.

Tobin, J. (1956). Estimation of Relationships for Limited Dependent Variables. *Econometrica*, **26**, 24-36.

Wooldridge J.M. (2003). *Econometrics Analysis of Cross Section and Panel data.* The MIT Press, Cambridge (MA).

Statistical Challenges in Modelling Operational Risk

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

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

Abstract: Using new tools, through the coefficient of variation, it is shown that alternative models for the classical generalized Pareto distribution may be more appropriate as statistical models in operational risk.

Keywords: Extreme Value Theory, coefficient of variation, heavy tails.

1 Introduction

Operational risk is defined, under the Basel II accord, as the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events. The Loss Distribution Approach (LDA) used to measure operational risk has three essential components: a distribution of the annual number of losses (frequency), a distribution of the *loss events* (known as the severity distribution) and an *aggregate loss* distribution that combines the two.
The severity distribution is not fixed by the Basel II capital accord. It's really a challenge for statistical methodology to model the severity distribution, since the classical approach that provides the Extreme Value Theory (EVT) does not seem appropriate. Although the generalized Pareto distribution (GPD) fits reasonable for most cases, it would not yield a reasonable capital estimate (sometimes more than 100% of the asset size), see Dutta and Perry (2007).

2 Dataset and Models

We will use the Danish data on fire insurance losses from January 1980 to December 1990 as a proxy of operational risk losses, following Degen et al. (2006).
This dataset has been considered in Resnick (1997) for assessing the appropriateness of heavy tailed models and testing for independence. In both cases there is no evidence to reject them. Embrechts et al. (1997) expose that there are difficult situations: to decide the threshold and the tail index parameter. Every author remarks the instability to evaluate the tail index. In general, these problems are common in operational risk.
A loss event L_i (also known as the loss severity) is an incident for which

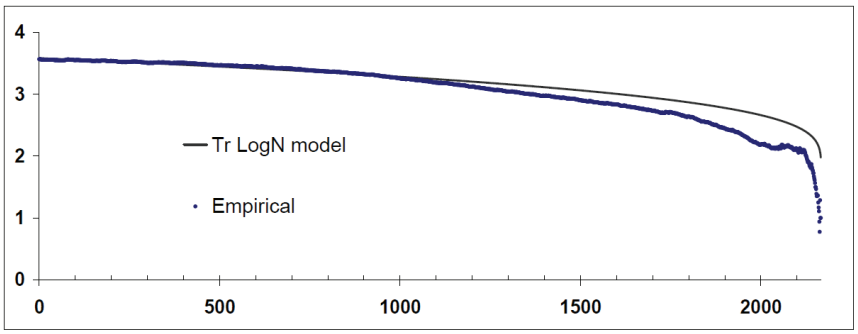


FIGURE 1. The theorical CV-plot for the TrLogN model using the whole sample fits very good to empirical data.

an entity suffers damages that can be measured with a monetary value. An aggregate loss over a specified period of time can be expressed as the sum

$$S = \sum_{i=1}^{N} L_i$$

where N is a random variable that represents the frequency of losses that occur over the period. We assume that the L_i are independent and identically distributed, and each L_i is independent from N.
Two models for severities are compared in this paper: GPD and truncated log-normal distribution (TrLogN). It is well known, from the theorem of Pickands-Balkema-de Haan (see Embrechts et al., 1997) that the conditional distribution of any random variable over a high threshold is approximately GPD. However, for practical applications, it has to be assumed that a threshold is taken sufficiently high and often this leads to limited data. Therefore, new models have to be considered.

3 Main results and Discussion

This paper shows the usefulness of the coefficient of variation in modeling the tails of the distributions. Moreover, beyond the VaR, new indicators to compare the models are considered. From Castillo and Daoudi (2009), it raises a new tool to deal with extreme values: the conditional coefficient of variation, considered as a random process in terms of the threshold. Let $\{x_k\}$ be a given sample of positive numbers of size n, we denote by CV-plot the representation of the sampling coefficient of variation (cv) of the conditional exceedance, given by

$$k \mapsto cv(\{x_{(j)} - x_{(k)} : j \geq k\})$$

TABLE 1. Quantile estimation (VaR) of Danish data with GPD and TrLogN models, using some threshold, u (sample of size 512, 254 and 109).

	Obs.	GPD model			TrLogN model		
		$u = 3.1$	$u = 5$	$u = 10$	$u = 3.1$	$u = 5$	$u = 10$
90%	5.5	5.7	5.6	5.5	5.7	5.6	5.5
95%	10.0	9.2	9.3	10.1	9.3	9.4	10.1
99%	26.0	27.5	27.5	27.3	27.1	27.5	27.4
99.9%	131.6	129.2	121.4	94.5	114.2	108.1	97.1

Figure 1 shows that empirical cv is not constant as it is in GPD, instead resembles the cv of the truncated log-normal fine. This is explained by the following theorem.

Theorem: The conditional CV of the log-normal distribution tends to 1 as the threshold goes to infinity.

Table 1 shows that the Peaks over Threshold method can be used with truncated log-normal as well as with GPD. Moreover, looking at the condicional value at risk (CVaR), in Table 2, we can see that the truncated log-normal model is more reasonable and stable than the GPD model.
The CV-plot in Figure 2 corresponds to the danish fire insurance data over threshold $u = 10$ transformed to exponentiality by the GPD and truncated log-normal models. It also shows that the use of the truncated log-normal model is more reasonable than the GPD model.

The good fit with the TrLogN model suggests the possibility that the tail index for GPD model can not be estimated, i.e. the domain of attraction may be the exponential model. Indeed, despite the theoretical convergence to 1, we see the slow convergence in Figure 1. Embrechts et al. (2007) prove that the log-normal distribution is subexponential but not a regular vari-

TABLE 2. CVaR estimation of Danish data with two models using some threshold, u (sample size 512, 254 and 109).

	Obs.	GPD model			TrLogN model		
		$u = 3.1$	$u = 5$	$u = 10$	$u = 3.1$	$u = 5$	$u = 10$
90%	15.6	17.7	17.1	15.8	16.2	15.9	15.9
95%	24.1	28.3	27.0	24.0	25.3	24.6	23.9
99%	58.6	83.5	76.6	58.3	68.7	64.1	58.4
99.9%	186.8	389.8	332.1	192.5	265.3	223.3	185.1

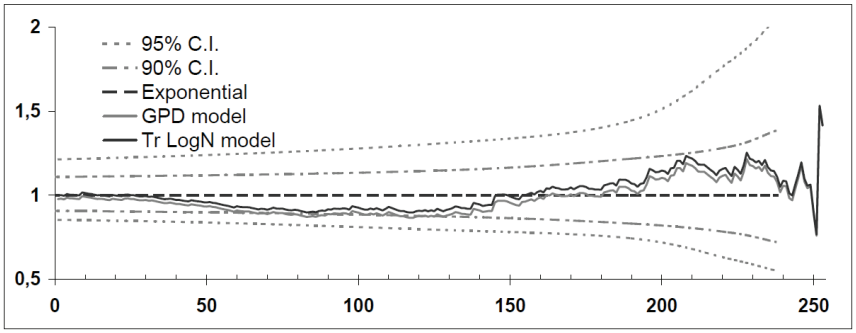


FIGURE 2. CV-plot of the trasformed to exponential tail (k=254) of Danish fire insurance data with GPD model and TrLogN model.

ation. This motivates us to look for models for the severity distribution of operational risk in this class. New models and methodology are nedeed, CV-plot approach allows to specify the model and simultaneously choose the threshold, as alternative to the Hill-plot. Finally, we remark that the Extreme Value Theory is appropiate but the main point is to search alternatives to GPD.
We conclude that the TrLogN model is more appropiate than the GPD model, for the dataset considered. The main reason perhaps is the slow convergence of the true model to the GPD.

References

Castillo, J. and Daoudi, J. (2009). Estimation of the generalized Pareto distribution. *Statistics and Probability Letters*, **79**, 684-688.

Degen, M., Embrechts, P. and Lambrigger, D. (2007). The Quantitative Modeling of Operational Risk: Betwen g-and-h and EVT. *Astin Bulletin*, **37**, 265-291.

Dutta, K. and Perry, J. (2006). A tale of tails: an empirical analysis of loss distribution models for estimating operational risk capital. *Working Paper. Federal Reserve Bank of Boston*, **06-13**.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events*. Springer-Verlag.

Resnick, S.I. (1997). Discussion of the Danish data on large fire insurance losses. *Astin Bulletin*, **27**, 139-151.

Is a Marginal Discrete Hidden Markov Model Lumpable?

Roberto Colombi[1], Sabrina Giordano[2]

[1] Università di Bergamo - Italy, colombi@unibg.it;
[2] Università della Calabria - Italy, sabrina.giordano@unical.it

Abstract: In this work, we suggest a way of avoiding the non-standard problem of testing lumpability of a Discrete Hidden Markov Model (DHMM) by treating the DHMM, whose lumpability is the focus, as a marginal process of a higher dimensional identifiable DHMM. The technique is then applied to real data.

Keywords: Markov Chains; Multinomial Hidden Markov Models; Conditional Independence; Identifiability

1 Introduction

A hidden Markov model is a versatile and computationally tractable tool for modelling time series widely applicable in fields such as Economics, Medicine, Biology and Engineering (MacDonald and Zucchini, 1997, Cappè et al., 2005).
In this work, we focus on discrete hidden Markov models (DHMMs) with a multivariate categorical observable process. One of the main problems related to hidden Markov models concerns the determination of an appropriate state space of the latent chain.This problem is approached from a perspective based on the lumpability property of Markov chains examined by Kemeny and Snell (1960) among others. These authors showed that when the state space cardinality of an observable Markov chain is reduced by collapsing certain states, the lumpability conditions need to be satisfied in order to preserve the Markovian property.
Here, we give a definition of a lumped version of a DHMM, according to which the lumped latent process is still a Markov chain with state space of smaller cardinality, and the observed variable depends on the latent process only through the aggregated latent states.
Unfortunately, testing the lumpability null hypothesis leads to a non-standard problem due to parameters which appear in the log-likelihood function only under the alternative hypothesis.
Our aim is to show that testing the lumpability conditions on a marginal DHMM of an identifiable multivariate DHMM does not suffer from the problem caused by non-identifiable parameters under the null hypothesis.

An example of the usefulness of testing lumpability of a marginal DHMM is given by the finance framework where the joint dynamics of m financial series, observed on an ordinal scale, such as rating asset series, may depend on the unobserved "Market states". The Market may switch from a state of low volatility to a state of high volatility through intermediate levels. Provided that a DHMM with r Market regimes can represent the dynamic pattern of the m financial series, one question is whether the dynamics of a marginal series is consistent with the hypothesis that the latent regimes are s lumped states of the original r ones, $s < r$. For instance, the DHMM for a multidimensional financial series may have $r = 5$ latent Market states of *very low, low, medium, high, very high* turbulence while for a marginal observed series, the coarser set of $s = 3$ Market states of *very low-low, medium, high-very high* turbulence may be appropriate.

2 Lumped discrete hidden Markov models

In this Section, we present the conditions to be satisfied for a DHMM to be lumpable.
Let $\{\boldsymbol{F}_t\}$ be an m-dimensional process $\{F_{1t}, F_{2t}, ..., F_{mt}\}$ whose marginal component $\{F_{jt}\}$ takes values in a finite set $\mathcal{F}_j$ with c_j categories, $j = 1, 2, ..., m$. One realization of this process at a given time is denoted by $\boldsymbol{f} = \{f_1, f_2, ..., f_m\} \in \mathcal{F}$, where $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2 \times ... \times \mathcal{F}_m$ has cardinality $c = \prod_{j=1}^{m} c_j$. The process $\{E_t, \boldsymbol{F}_t\}$ is a DHHM if it is a first order multivariate Markov chain, $\{E_t\}$ is a non-observable process with finite state space $\mathcal{E} = \{e_i, i = 1, ..., r\}$ and the following two conditions hold

C1) $E_t \perp\!\!\!\perp \boldsymbol{F}_{t-1} | E_{t-1}$,

C2) $\boldsymbol{F}_t \perp\!\!\!\perp E_{t-1}\boldsymbol{F}_{t-1} | E_t$.

The previous assumptions imply that $\{E_t\}$ is a first order Markov chain and the marginal processes $\{E_t, \boldsymbol{F}_{jt}\}$, $j = 1, 2, ..., m$ are DHMMs as well.
Let $(g, \boldsymbol{f}) = (h(e), \boldsymbol{f})$, $(e, \boldsymbol{f}) \in \mathcal{E} \times \mathcal{F}$, be a function that maps the state space $\mathcal{E} \times \mathcal{F} = \{(e_i, \boldsymbol{f}), i = 1, ..., r, \boldsymbol{f} \in \mathcal{F}\}$ of a DHMM $\{E_t, \boldsymbol{F}_t\}$ onto a set $\mathcal{G} \times \mathcal{F} = \{(g_k, \boldsymbol{f}), k = 1, ..., s, \boldsymbol{f} \in \mathcal{F}\}$, where s is smaller than r. Then $\{G_t, \boldsymbol{F}_t\}$, so that $(G_t, \boldsymbol{F}_t) = (h(E_t), \boldsymbol{F}_t)$, is a *lumped version of the DHMM* $\{E_t, \boldsymbol{F}_t\}$ if the following two conditions are satisfied

C3) $G_t \perp\!\!\!\perp E_{t-1} | G_{t-1}$,

C4) $\boldsymbol{F}_t \perp\!\!\!\perp E_t | G_t$.

We shall also refer to a lumped state $g \in \mathcal{G}$ by the term *macrostate* since it is created by aggregating original single states $e \in \mathcal{E}$, called *microstates*. Condition *C3* ensures that the marginal lumped process, obtained by aggregating the states of the latent component and marginalizing the DHMM with respect to the observable process, is a Markov chain. Condition *C4*,

on the other hand, means that the observable process does not depend on the original hidden microstate at the present time given the current macrostate. This condition seems a key feature, in fact, when the aim is to lump states together, it is essential that all the states to be aggregated provide the same information.

3 Testing lumpability

In this Section we first clarify why testing the lumpability of a DHMM $\{E_t, \boldsymbol{F}_t\}$ is a non-standard problem due to unrestricted parameters which are not identifiable under the null hypothesis, and then we illustrate that the testing of lumpability conditions on a marginal DHMM of $\{E_t, \boldsymbol{F}_t\}$ does not encounter the same identifiability problem.

3.1 A non-standard hypothesis testing problem for a DHMM

We begin by introducing useful notations. The transition probabilities of $\{E_t\}$ from a state $e \in \mathcal{E}$ to a state $e' \in \mathcal{E}$ are denoted by $q(e', e)$, and $\psi(g', e) = \sum_{e':h(e')=g'} q(e', e)$ stand for the transition probabilities from one of the microstates $e \in \mathcal{E}$ to a lumped state $g' \in \mathcal{G}$.
For a DHMM, condition *C3* is equivalent to the constraints

$$\psi(g', e) = \tau(g', g), \quad \forall e : h(e) = g, \quad g', g \in \mathcal{G}. \tag{1}$$

There are $r(s-1)$ linearly independent probabilities $\psi(g', e)$, while the independent $\tau(g', g)$ are $s(s-1)$, and consequently $(r-s)(s-1)$ independent constraints are to be required in order to satisfy the equalities (1).
The joint probability function of $\boldsymbol{F}_t$ given $E_t = e$ is assumed to be time invariant and is denoted by $p(\boldsymbol{f}, e)$, while $k(\boldsymbol{f}, g')$ represents the probability of $\boldsymbol{F}_t = \boldsymbol{f}$ given the latent macrostate $G_t = g'$.
There are $r(c-1)$ probabilities $p(\boldsymbol{f}, e)$ and $s(c-1)$ probabilities $k(\boldsymbol{f}, g')$, thus condition *C4* is equivalent to the $(r-s)(c-1)$ independent constraints

$$p(\boldsymbol{f}, e') = k(\boldsymbol{f}, g'), \quad \forall e' : h(e') = g' \quad g' \in \mathcal{G}. \tag{2}$$

Testing (1) and (2) is a non-standard problem. This is due to the fact that the log-likelihood function under the $(r-s)(c+s-2)$ lumpability constraints depends only on probabilities $\tau(g', g)$ and $k(\boldsymbol{f}, g')$, whereas the $r(r-s)$ unrestricted parameters $\varphi(e'|g', e) = \frac{q(e', e)}{\psi(g', e)}$, which denote the transition probabilities from $e \in \mathcal{E}$ to $e' \in \mathcal{E}$, given that $h(e') = g'$, $g' \in \mathcal{G}$, are not present in the log-likelihood function under the null hypothesis. To prove this statement, $\boldsymbol{f}_t$ will denote the observed m-dimensional response at time t and, e_t and g_t will specify the hidden microstate and macrostate at time t, respectively, in the remaining part of this Section only.

From (1) and (2), for every e_{t-1} so that $h(e_{t-1}) = g_{t-1}$, $t = 1, 2, ..., T$, it holds

$$\sum_{e_t} q(e_t, e_{t-1}) p(\boldsymbol{f}_t, e_t) = \sum_{g_t} \sum_{e_t : h(e_t) = g_t} \psi(g_t, e_{t-1}) \varphi(e_t | g_t, e_{t-1}) p(\boldsymbol{f}_t, e_t) =$$

$$= \sum_{g_t} \tau(g_t, g_{t-1}) \sum_{e_t : h(e_t) = g_t} \varphi(e_t | g_t, e_{t-1}) p(\boldsymbol{f}_t, e_t) = \sum_{g_t} \tau(g_t, g_{t-1}) k(\boldsymbol{f}_t, g_t).$$

By applying the previous equality for $t = T, T-1, ..., 1$ recursively, we obtain the likelihood function under constraints (1) and (2)

$$\begin{aligned} l(\boldsymbol{f}_1, \boldsymbol{f}_2, ..., \boldsymbol{f}_T) &= \sum_{e_0} \pi(e_0) \sum_{\mathbf{e}_t} \prod_{t=1}^{T} \psi(g_t, e_{t-1}) \varphi(e_t | g_t, e_{t-1}) p(\boldsymbol{f}_t, e_t) = \\ &= \sum_{g_0} \sum_{e_0 : h(e_0) = g_0} \pi(e_0) \sum_{\mathbf{g}_t} \prod_{t=1}^{T} \tau(g_t, g_{t-1}) k(\boldsymbol{f}_t, g_t). \end{aligned} \tag{3}$$

where $\pi(e_0)$ are the probabilities of the initial distribution of the latent chain, and $\sum_{\mathbf{e}_t}$, $\sum_{\mathbf{g}_t}$ denote the sums with respect to all realizations of the latent process and the lumped one, respectively.
Equation (3) shows that the unrestricted parameters $\varphi(e'|g', e)$ do not appear in the likelihood function under the null hypothesis of lumpability.
Since a number of parameters are not identifiable under the null hypothesis, the asymptotic distribution of the LRT statistic is no longer chi-square and, therefore, testing *C3* and *C4* gives rise to a non-standard problem, (Andrews, 2001).

3.2 A standard testing problem for a marginal DHMM

We now turn our attention to proving that the difficulties due to non-identifiable parameters under the null hypothesis do not arise when we test the lumpability conditions on a marginal process of an identifiable multivariate DHMM. It is worth noting that when a DHMM is identifiable it is no longer lumpable, as otherwise the lumpability conditions would lead to the aforementioned identifiability problems.
Let us partition the observable process of an m-dimensional identifiable DHMM $\{E_t, \boldsymbol{F}_t\}$ into two marginal processes $\{\boldsymbol{F}_t\} = \{\boldsymbol{F}_{1t}, \boldsymbol{F}_{2t}\}$. To see if $\{E_t, \boldsymbol{F}_{1t}\}$ is lumpable we need to test the restrictions (1) on the transition probabilities and the conditions

$$p_1(\boldsymbol{f}_1, e') = k_1(\boldsymbol{f}_1, g'), \quad \forall e' : h(e') = g', \quad g' \in \mathcal{G} \tag{4}$$

on the marginal distribution of $\boldsymbol{F}_{1t}$ given E_t. The lumpability restrictions (1) and (4) for the marginal DHMM require a number of $(r-s)(c_1+s-2)$ constraints as shown before, but the identifiability problem of the previous Section does not occur. In fact, now the joint log-likelihood function

depends on all parameters as the joint distributions $p(\boldsymbol{f}, e')$ cannot fulfil condition (2) because otherwise this condition, together with (1), would contradict the assumed hypothesis that the joint DHMM is identifiable. Therefore, the usual chi-square LRT statistic applies for testing equalities (1) and (4) and then the lumpability property for the marginal DHMM is ascertained through a standard hypothesis testing procedure.
Thus, if we want to test the hypothesis that $\{E_t, \boldsymbol{F}_{1t}\}$ is lumpable, without incurring a non-standard problem, it is enough to find an auxiliary observable process $\{\boldsymbol{F}_{2t}\}$ so that $\{E_t, \boldsymbol{F}_{1t}, \boldsymbol{F}_{2t}\}$ is an identifiable DHMM.

4 Example

This Section exemplifies the problem of testing lumpability firstly on a categorical time series of average weekly sales quantities of two kinds of *cheese* (*spread* and *fresh*), sold in 2221 Italian shopping centres from January 1, 2000 to December 31, 2004. The weekly sales averages were classified into

TABLE 1. Tests of lumpability in DHMM for spread cheese data

s	*lumped states*	*LRT*	*df*	*p-value*
2	(e_1, e_2) (e_3, e_4)	34.52	6	0.0000
2	(e_1) (e_2, e_3, e_4)	29.68	6	0.0000
2	(e_1, e_2, e_3) (e_4)	28.41	6	0.0000
2	(e_1, e_4) (e_2, e_3)	24.33	6	0.0004
2	(e_1, e_3) (e_2, e_4)	24.32	6	0.0004
2	(e_3) (e_1, e_2, e_4)	19.50	6	0.0034
2	(e_2) (e_1, e_3, e_4)	20.60	6	0.0021
3	$(e_1)(e_2, e_3)$ (e_4)	11.74	4	0.0193
3	(e_1, e_2) $(e_3)(e_4)$	10.36	4	0.0347
3	$(e_1)(e_2)$ (e_3, e_4)	6.661	4	0.1549

3 conventional levels *low, medium, high* according to the brand specifications. We comment on a few results of lumpability hypotheses testing for the sequence of the *spread cheese* sale levels, when the *fresh cheese* sales series is the auxiliary variable. Here, the reference DHMM for the bivariate series of *spread* and *fresh cheese* is assumed to have 4 latent states. Table 1 illustrates the results of testing certain lumpability hypotheses. A lumped version with 3 macrostates fits the marginal series of the *spread cheese* sales levels. In fact, the possibility to reduce the number of latent states to 3 is undoubtedly accepted when e_3 and e_4 are collapsed. The hypotheses of lumping the 4 hidden states in any other 3 sets are all rejected instead and there is no evidence to maintain that a 2 states lumped version of the

TABLE 2. Tests of lumpability in DHMM for soft drink data

lumped states	*LRT*	*df*	*p-value*
(e_1, e_2) (e_3, e_4)	8.902	6	0.1791
(e_1, e_4) (e_2, e_3)	5.839	6	0.4413
(e_3) (e_1, e_2, e_4)	12.47	6	0.0522
(e_2) (e_1, e_3, e_4)	5.851	6	0.4401

original DHMM may fit the *spread cheese* sales series, as summarized in Table 1. In the second example we use the data of a soft drink company (Ching et al., 2002). The sequence of data, here analyzed, consists of a one-year time series of daily sales of soft drinks: *mint tea*, *lemon tea* with categories: *low, medium, high* level. If the *lemon tea* sales series is the relevant variable, the hypothesis of lumping the 4 initial microstates into 2 macrostates can be accepted when the states are collapsed in 4 different ways (Table 2). All the hypotheses were tested using the R-package *hmmm*, Colombi and Cazzaro (2008).

References

Andrews, D.W.K. (2001). Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica*, **69**, 683-734.

Cappè, O., Moulines, E., Rydén T. (2005). *Inference in Hidden Markov Models.* New York: Springer.

Ching, W.K., Fung, E.S., Ng, M.K. (2002). A multivariate Markov chain for categorical data sequences and its applications in demand predictions. *IMA Journal of Management Mathemathics*, **13**, 187-199.

Colombi, R., Cazzaro, M. (2008). Hierarchical multinomial marginal models: the R package *hmmm*. *www.unibg.it/pers/?colombi*.

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

Kemeny, J.C., Snell, J.L. (1960). *Finite Markov Chains*. Princeton: D.Van Nostrand.

Mean-Variability Hidden Markov Models for the detection of influenza outbreaks.

David Conesa[1], Rubén Amorós[1], Antonio López-Quílez[1], Miguel A. Martínez-Beneito[2]

[1] Department of Statistics and Operations Research, Universitat de València, C/ Dr. Moliner 50, 46100 Burjassot (Valencia), Spain (correspondence address David.V.Conesa at uv.es)

[2] Centro Superior de Investigación en Salud Pública, Generalitat Valenciana

Abstract: Considerable effort has been devoted to the development of statistical algorithms for the automated monitoring of influenza surveillance data. In this work, we introduce a framework of models in order to early detect the onset of an influenza epidemic based on different kind of surveillance data. In particular, the process of the observed cases is modelled via a Bayesian Hierarchical Poisson model in which the intensity parameter is a function of the incidence rate. The key point is to consider this incidence rate as a normal distribution in which both parameters are modelled differently depending if the system is in epidemic or non epidemic phase. To do so, we use a hidden Markov model in which transition between both phases is modelled as a Markovian process. Different options for modelling the mean of the rates are described, including the option of modelling the mean at each phase as autoregressive processes of order 0, 1 and 2. Bayesian inference is carried out to provide the probability of being in an epidemic state at any given moment. Methodology is applied on various influenza illness data-sets. Results indicate that our methods perform better than previous approaches in terms of sensitivity, specificity and timeliness.

Keywords: Autoregressive modelling; Bayesian inference; Hidden Markov models; Public health; Temporal Surveillance.

1 Introduction

Public Health agencies use disease surveillance tools in order to monitor the incidence or prevalence of specific health problems over time. This knowledge allows them to detect changes in the estimated incidence rates, which produces better planning and allocation of resources and the possibility of avoiding breakdowns in Health Care Systems. In addition, a good surveillance infrastructure can be very useful in preparing for pandemics and for monitoring new emerging diseases.

Among other diseases, influenza has been of special interest among researchers as influenza epidemics occur virtually every year and result in

substantial disease, death and expense. As a result, a large variety of statistical algorithms for the automated monitoring of influenza surveillance have been proposed.
Martínez-Beneito et al. (2008) proposed a Markov switching model to determine the epidemic and non-epidemic periods in an influenza season using the series of differenced incidence rates. In a subsequent paper, Conesa et al. (2009) introduced `FluDetWeb`, an implementation of Martínez-Beneito et al.'s model based on a client-server architecture with a thin (web-based) client application design.
The main purpose of this work is to present an enhanced modelling of Martínez-Beneito et al. (2008). In our proposal we take profit of all advantages of Martínez-Beneito et al. model (use of Markov models and variability of data in order to distinguish between both epidemic and non-epidemic phases). But we also incorporate the magnitude of the incidence of the disease in the model (low incidence meaning clearly non-epidemic phase). The last novelty relies on the fact that we model the obtained cases instead of the incidence rates.

2 The model

Let $O_{i,j}$ denote the observed cases of influenza of week i in year j. Our model is based on:

$$\begin{aligned} O_{i,j} &\sim \text{Poisson}(\nu_{i,j}) \\ \nu_{i,j} &= f(r_{i,j}) \\ r_{i,j} &\sim \mathcal{N}(R_{i,j}(Z_{i,j}), \sigma_j^2(Z_{i,j})) \end{aligned} \tag{1}$$

The function in the second line depends on the type of data we are working with. In the usual situation in which we want to model standardized rates per 100.000 inhabitants, this expression would be:

$$\nu_{i,j} = \frac{Pop_{i,j} \cdot r_{i,j}}{100.000}\,, \tag{2}$$

where $Pop_{i,j}$ represents the population from where the observed cases have been reported on the corresponding week (or day).
It is worth mentioning that we model the rate as a normal distribution in which both the mean and variance depend on $Z_{i,j}$, an unobserved random variable that indicates which phase the system is in (1, epidemic; 0, endemic). This is the basic idea of a Markov switching models in which the unobserved sequence of $Z_{i,j}$ (note that we do not know which phase the system is in at any given moment) follows a two-state Markov chain of order 1 with transition probabilities:

$$P(Z_{i+1,j} = l | Z_{i,j} = k) = P_{k,l} \;\; k, l \in \{0, 1\} \tag{3}$$

At the next stage of the hierarchy, we have to model the mean and the variance of the rates in both phases. With respect to the variance, we assume constant but different variances for each phase of each season. We also assume a lower variance in the endemic phase. Including the variability of the observed data in the model has proved to be very helpful to distinguish between both epidemic and non-epidemic phases, as non-epidemic dynamics are characterized by small random changes, while in the epidemic dynamics changes are greater.
Next step is to model the mean of the rates in both states. Note that importance of $R_{i,j}(0)$ and $R_{i,j}(1)$ comes from the fact that they represent the magnitude of the incidence at each phase and so we can take profit of it to distinguish between both dynamics. Different options for modelling the mean of the rates are described, including the option of modelling the mean at each phase as autoregressive processes of order 0, 1 and 2. For example, when modelling them with autoregressive processes of order 1:

$$\begin{aligned} R_{i,j}(0) &= \mu_0 + \rho_0 \cdot (r_{i-1,j} - \mu_0) \\ R_{i,j}(1) &= \mu_1 + \rho_1 \cdot (r_{i-1,j} - \mu_1) \\ \mu_0 &< \mu_1 \end{aligned}$$

Once the model is determined, the following step is to estimate its parameters. We use the Bayesian paradigm, which requires specification of the prior distributions of each parameter involved in the model:

$$\begin{aligned} P_{1,1} &\sim Beta(0.5, 0.5) \\ P_{0,0} &\sim Beta(0.5, 0.5) \\ \sigma_j(0) &\sim Unif(\theta_{low}, \theta_{mid1}) \\ \sigma_j(1) &\sim Unif(\theta_{mid2}, \theta_{up}) \\ \theta_{low} &= \lambda_{[1]} \\ \theta_{mid1} &= \lambda_{[2]} \\ \theta_{mid2} &= \lambda_{[3]} \\ \theta_{up} &= \lambda_{[4]} \\ \lambda_j &\sim Unif(a, b) \;\; j = 1, ..., 4 \\ \mu_0 &= \theta_{[1]} \\ \mu_1 &= \theta_{[2]} \\ \theta_j &\sim Unif(0, +\infty) \;\; j = 1, 2 \\ \rho_0, \rho_1 &\sim Unif(-1, 1)\,. \end{aligned} \tag{4}$$

3 Results and discussion

We finally present an application of our modelling on various influenza illness data-sets, in particular, data obtained from the North Carolina influenza Sentinel surveillance program and also from the Google Flu Trends

database. When possible, we also perform a quantitative study of the detection power of our method. The performance of an early warning system can be measured by its sensitivity, specificity and timeliness. Cowling et al. (2006) and Kleinmann and Abrams (2006) have proposed different metrics that combine this three characteristics in order to provide an overall measure of performance. In particular, we use AUWROC1, VUTROS1, VUTROS3 and VUTROCS (see both papers for more details). Results indicate that our methods perform better than previous approaches in terms of sensitivity, specificity and timeliness.

Acknowledgments: Financial support from the Conselleria de Sanitat of the Generalitat Valenciana (the Valencian Regional Health Authority) is gratefully acknowledged. The authors would also like to acknowledge financial support from the Spanish Ministry of Education and Science via research grants MTM2007-61554 (jointly financed with the European Regional Development Fund) and FUT-C2-0047 (as part of the INGENIO-MATHEMATICA research project) and from the Generalitat Valenciana via research grants GV/2007/079, AP-049/08 and EVES-015/2008.

References

Burkom, H. (2007). Alerting algorithms for Biosurveillance. In: *Disease surveillance, a public health informatics approach*. Edited by Lombardo J.S. and Buckeridge D.L. John Wiley and Sons, Ltd. 143–192.

Conesa, D., López-Quílez, A., Martínez-Beneito, M.A., et al. (2009). FluDetWeb: an interactive web-based system for the early detection of the onset of influenza epidemics. *BMC Medical Informatics and Decision Making*, **9**:36.

Cowling, B.J., Wong, I.O.L., Riley, S. and Leung, B. M. (2006). Methods for monitoring influenza surveillance data. *International Journal of Epidemiology*, vol. 35, pp. 1314–1321.

Kleinman, K.P., Abrams, A. M. (2006). Assessing surveillance using sensitivity, specificity and timeliness. *Statistical Methods in Medical Research*, vol. 15, pp. 445–464.

LeStrat, Y. (2005). Overview of temporal surveillance. In: *Spatial and syndromic surveillance for public health*. Edited by Lawson, A.B. and Kleinman, K. John Wiley and Sons, Ltd. 13–29.

Martínez-Beneito, M. A., Conesa, D., López-Quílez, A., López-Maside, A. (2008). Bayesian Markov switching model s for the early detection of influenza epidemics. *Statistics in Medicine*, **27**(22) 4455–4468.

Bartlett Correction in Power Series Generalized Nonlinear Models

Audrey H.M.A. Cysneiros[1], Gauss M. Cordeiro[2], Priscila Gonçalves da Silva[3]

[1] Universidade Federal de Pernambuco - Brazil, e-mail: audrey@de.ufpe.br
[2] Universidade Federal de Pernambuco - Brazil, e-mail: priscilaestat@yahoo.com.br
[3] Universidade Federal Rural de Pernambuco - Brazil, e-mail: cordeiro@deinfo.ufrpe.br

Abstract: This paper obtains Bartlett correction to the likelihood ratio statistic in Power Series generalized Nonlinear Models, considering the dispersion parameter known. We have numerically evaluated the finite sample performance of likelihood ratio tests and its Bartlett-corrected versions on the size and power.

Keywords: Bartlett correction; Generalized linear models; Likelihood ratio test; Maximum likelihood

1 Introduction

In the power series generalized nonlinear models, we have the possibility of a more general family of discrete distributions for the response variable and a nonlinear structure for the regression parameters, although the dispersion parameter is assumed to be a constant and other shape parameters are assumed to be known exactly. In this paper we obtain Bartlett correction to the likelihood ratio statistic in this class of models. Numerical results are presented.

2 Power Series Generalized Nonlinear Models

We consider discrete random variables $Y_1, \cdots, Y_n$ in Y which are independent and each Y_i follows a family of distributions with mean parameter $\mu_i > 0$ and dispersion parameter $\phi > 0$ defined by the probability mass function with respect to Lebesgue measure

$$\pi(y; \mu_i, \phi) = \frac{a(y, \phi) g(\mu_i, \phi)^y}{f(\mu_i, \phi)}, \qquad y \in A_s, \tag{1}$$

where the support of Y_i is a subset A_s of integers $\{s, s+1; \cdots\}$ defined here not depending upon unknown parameters, $s \geq 0$, $a(y; \phi)$ is positive, and the analytic functions $f(\mu_i; \phi)$ and $g(\mu_i; \phi)$ (of the mean parameter

μ_i and the common dispersion parameter ϕ) are positive, finite and twice-differentiable. Here, the dispersion parameter ϕ is assumed known, ($\phi > 0$). We have $E(Y) = \mu = \frac{f'g}{fg'}$ and $Var(Y) = V(\mu, \phi) = \frac{g}{g'}$, with $f = f(\mu_i, \phi)$ and $g = g(\mu_i, \phi)$ where from now on the primes denote differentiation with respect to μ. We introduce a nonlinear regression structure for the class of distributions (1) through a systematic component for the mean vector $\mu = E(Y_i)$ given by $h(\mu_i) = \eta_i = \eta(x_i; \beta)$ where $h(\cdot)$ is a known one-to-one differentiable link function, $\eta(\cdot; \cdot)$ is a specified nonlinear function of unknown parameters, x_i is a $q \times 1$ vector and $\beta = (\beta_1, \cdots, \beta_p)^\top$ for $p < n$ is a set of unknown parameters to be estimated. Further, we assume that β is defined in a subset Ω_β of $I\!R^p$ ($p < n$) and $\eta(x_i; \beta)$ is an injective and continuously differentiable function with respect to β such that the $n \times p$ derivative matrix of the nonlinear predictor, $\tilde{X} = \tilde{X}(\beta) = \partial\eta/\partial\beta^\top$ say, has rank p for all β. The $n \times p$ local model matrix $\tilde{X}$ in general depends on the unknown parameter β.

3 Improved likelihood ratio test

The asymptotic chi-squared distribution of the likelihood ratio (LR) statistic is frequently used to test hypotheses of interest in regression models. However, for small sample size n, the use of such statistic has less justification. An alternative is to use a higher order asymptotic theory. Bartlett (1937) proposed an improved LR statistic. His argument goes as follows. Suppose that under the null hypothesis $E(LR) = q\{1+d+O(n^{-2})\}$, where d is a constant that can be consistently estimated under the null hypothesis and q is the difference of the dimensions of the parameter spaces under the alternative and null hypotheses. Then, the expected value of the transformed statistic $LR^* = LR/(1+d)$ and $LR_1^* = LR(1-d)$ is closer to the one from χ^2_q distribution than the expected value of LR. This became widely known as the Bartlett correction. LR statistic for nonlinear hypotheses on the regression coefficients can have sizes that are typically larger than their nominal sizes. Then, the Bartlett correction, under mild regularity conditions, guarantees that all the moments of the adjusted LR^* statistic are equal to those of the asymptotic χ^2_q distribution up to order n^{-1}. Thus, in matrix notation we obtain the Bartlett correction as $d = \epsilon_p - \epsilon_{p-q}/q$ where $\epsilon_p = \epsilon_p^{(L)} + \epsilon_p^{(NL)}$ with $\epsilon_p^{(L)} = \iota^\top Z_d Q_4 Z_d \iota + \iota^\top Q_1 Z^{(3)} Q_2 \iota + \iota^\top \tilde{W}_1 Z^{(3)} \tilde{W}_1 \iota + \iota^\top Q_1 Z_d Z Z_d Q_3 \iota + \iota^\top \tilde{W}_1 Z_d Z Z_d \tilde{W}_1 \iota$, and $\epsilon_p^{(NL)} = \iota^\top \Big(\tilde{W}_1 - \frac{1}{2} Q_1\Big) D Z_d \iota + \iota^\top (\tilde{W}_1 - Q_1) C_d \iota - \frac{1}{4} \iota^\top W_1 (2B - D^2) \iota + \frac{1}{4} \iota^\top D W_1 Z \Big[W_1 D + 4 Z_d \Big(\tilde{W}_1 - \frac{1}{2} Q_1\Big)\Big] \iota + tr\Big\{\Big[\Big(\tilde{W}_1 - Q_1\Big) C - \frac{1}{2} W_1 B\Big] W_1 Z\Big\}$, where $Z = \tilde{X}(\tilde{X}^\top W \tilde{X})^{-1} \tilde{X}^\top$, B and C are $n \times n$ matrices with (i,j)–elements defined by $b_{ij} = tr(K_\beta^{-1} \tilde{X}_i^* K_\beta^{-1} \tilde{X}_j^*)$ and $c_{ij} = \tilde{x}_i^\top K_\beta^{-1} \tilde{X}_j^* K_\beta^{-1} \tilde{x}_i^\top$, respectively. Also D =diag$\{d_{11}, \ldots, d_{1n}\}$ is a diagonal matrix with elements given by $d_{1i} = tr(K_\beta^{-1} \tilde{X}_i^*)$, $\tilde{W}_1$, W_1, Q_1, Q_2, Q_3 and Q_4 are

$n \times n$ matrices with (i,j)–elements defined by $w_{ji} = \left(\frac{f_i' g_i}{f_i g_i'} t_i^{(j)} - q_i^{(j)}\right) \frac{1}{h_i'}$, $\tilde{w}_{ji} = \varphi_{ji} - \frac{(j-1) q_i V_i t_i^{(j)} h_i'' - q_i^{(j+1)}}{(h_i')^{j+1}} + j \frac{q_i^{(j)} h_i''}{(h_i')^{j+2}}$, $\varphi_{ji} = \frac{q_i' V_i t_i^{(j)} + q_i V_i' t_i^{(j)} + q_i V_i t_i^{(j+1)}}{(h_i')^j}$, $q_{1i} = w_{2i} - \frac{w_{1i} h_i''}{(h_i')^2}$, $q_{2i} = \frac{1}{6}\left(w_{2i} - \frac{w_{1i} h_i''}{(h_i')^2}\right) - \tilde{w}_{1i}$, $q_{3i} = \frac{1}{4}\left(w_{2i} - \frac{w_{1i} h_i''}{(h_i')^2}\right) - \tilde{w}_{1i}$, $q_{4i} = \frac{1}{4} w_{3i} - \frac{3}{4}\frac{w_{2i} h_i''}{(h_i')^2} + \frac{3}{4}\frac{w_{1i} h_i'''}{(h_i')^3} - \frac{5}{4}\frac{w_{1i}(h_i'')^2}{(h_i')^4} + \frac{\tilde{w}_{1i} h_i''}{(h_i')^2} - \tilde{w}_{2i} + \tilde{w}_{1i}^*$, respectively, $i = 1, \cdots, n$, $j = 1, 2$ and ι is an $n \times 1$ vector of ones. Z_d and C_d represents the diagonal matrices obtained from the diagonal elements of Z and C.

4 Numerical Evidence

In Table 1 we report some simulation results in order to compare the sizes of the usual likelihood ratio test and of the tests based on the following modified likelihood ratio statistics: LR^* and LR_1^*. We use the following nonlinear regression model that assumes the predictor: $\eta_i = \beta_0 + \sum_{j=1}^{7} \beta_j x_{ij} + \exp(\beta_8 x_{8i}), i = 1, \cdots, n$. The null hypothesis that we consider is $\beta_5 = \beta_6 = 0$ and the response was generated from a Consul distribution, Generalized negative binomial (GNB) and from a Generalized Poisson (GP) distribution. Ten thousand samples of 30 observations were generated for each model with $p = 8$ (the number of regression parameters), $\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_7 = \beta_8 = 0.05$, $\beta_2 = 1$, ($\phi = 1$ – Consul), ($\phi = 1, \nu = 3$ – GNB), ($\phi = 0.2$ – GP), and the independent variable $x_1, x_2, x_3, x_4, x_5, x_6, x_7$ and x_8 were chosen as independent random draws from the following distributions: U(0,1), F(2,5), Cauchy, N(0,1), t_3, LN(0,1), χ_3^2 and F(3,3). Table 1 displays the null rejection rates of the three tests for 10 and 5% nominal levels (α). The figures in Table 1 reveal important information. It is clear that the size performance of the usual likelihood ratio test deteriorates as the number of nuisance of regression parameters increases. Second, all corrected likelihood ratio tests have simulated sizes closer to the nominal levels than the unmodified likelihood ratio test. Table 2 displays the null rejection rates for the sample sizes $n = 20, 30, 40, 50$ and $p = 8$. The likelihood ratio test is largely liberal for small samples, overrejecting the null hypothesis more frequently than expected based on the selected nominal levels. And, the corrections are really necessary for small and moderate sample sizes.

Acknowledgments: This study was supported by CNPq and FACEPE, Brazil.

References

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society A*, **160**, 268-282.

TABLE 1. Size simulations: rejection rates of the likelihood ratio and two corrected likelihood ratio tests, $n = 30$; entries are percentages

		GNB			Consul			GP		
$\alpha(\%)$	p	LR	LR^*	LR_1^*	LR	LR^*	LR_1^*	LR	LR^*	LR_1^*
5	3	5.8	5.1	5.1	5.8	5.0	5.0	5.3	5.0	5.0
	4	5.5	5.1	5.1	5.5	4.9	4.8	5.0	4.6	4.6
	5	5.7	5.3	5.3	5.8	5.0	5.0	5.3	5.1	5.1
	6	5.4	4.9	4.8	5.9	4.9	4.8	5.3	5.0	5.0
	7	6.5	5.1	5.1	7.0	4.9	5.0	6.1	4.8	4.8
	8	6.4	5.1	5.0	7.1	4.8	4.8	6.3	5.0	5.0
	9	6.7	5.0	5.0	7.9	5.0	4.6	6.5	5.0	5.0
10	3	10.6	10.2	10.2	10.5	10.2	10.2	10.2	10.0	10.0
	4	10.5	10.1	10.1	10.7	10.2	10.2	10.3	10.0	10.0
	5	11.0	10.2	10.2	11.6	10.4	10.4	11.1	10.4	10.4
	6	12.0	10.9	10.8	12.4	10.7	10.5	12.0	10.8	10.8
	7	12.8	10.8	10.6	13.7	10.9	10.5	12.7	10.9	10.7
	8	13.3	10.9	10.5	14.3	10.5	9.7	13.3	10.8	10.5
	9	14.9	11.1	10.3	16.8	10.9	9.7	14.6	11.0	10.1

TABLE 2. Size simulations: rejection rates of the likelihood ratio and two corrected likelihood ratio tests, $p = 8$; entries are percentages

		GNB			Consul			GP		
n	α	LR	LR^*	LR_1^*	LR	LR^*	LR_1^*	LR	LR^*	LR_1^*
20	1	2.5	1.3	1.3	2.7	1.1	2.0	2.5	1.2	1.4
	5	9.0	5.7	5.3	10.9	5.1	5.4	9.0	5.8	5.4
	10	15.4	10.8	9.8	17.8	9.9	9.4	15.8	10.8	10.0
30	1	1.8	1.2	1.2	2.3	1.4	1.2	1.8	1.2	1.2
	5	7.5	5.7	5.5	8.1	5.7	5.1	7.5	5.6	5.3
	10	13.3	10.9	10.5	14.3	10.5	9.7	13.3	10.8	10.5
40	1	1.6	1.2	1.2	1.7	1.2	1.1	1.4	1.2	1.1
	5	5.9	4.9	4.8	6.9	5.2	5.0	6.1	5.0	4.9
	10	11.6	9.8	9.6	12.6	10.1	9.7	11.5	10.0	9.8
50	1	1.3	1.0	1.0	1.5	1.1	1.1	1.3	1.0	1.0
	5	5.6	4.8	4.8	6.1	5.1	5.0	5.7	5.0	5.0
	10	11.1	9.9	9.8	11.5	9.9	9.8	10.9	9.7	9.6

Cordeiro, G. M., Andrade, M. G., De Castro, M. (2009). Power series generalized nonlinear models. *Computational Statistics and Data Analysis*, **53**, 1155-1166.

Birnbaum-Saunders Linear Regression Models: A New Approach

Francisco José A. Cysneiros[1], Víctor Leiva[2], Manoel Ferreira Santos-Neto[1]

[1] Departamento de Estatística, Universidade Federal de Pernambuco, Recife, PE, Brazil, e-mail: cysneiros@de.ufpe.br and mn.neco@gmail.com
[2] Departamento de Estadística, CIMFAV, Universidad de Valparaíso, Valparaíso, Chile, e-mail: victor.leiva@uv.cl; victor.leiva@yahoo.com

Abstract: Regression models based on the Birnbaum-Saunders (BS) distribution have been applied with success in different areas. In this work, we propose a new reparameterization of the BS distribution based on two parameters, mean and dispersion. From this new approach, we develop a model for the mean response variable by using a link function such as occurs in generalized linear models. In addition, we carry out a residual analysis and influence diagnostics for this model. Finally, we illustrate the proposed methodology by means of an example.

Keywords: Birnbaum-Saunders distributions; Local influence; Reparameterization; Residuals.

1 Birnbaum-Saunders linear regression models

In the last two decades, a positively skewed probability model with non-negative support that has received a great attention is the Birnbaum-Saunders (BS) distribution; for more details, see Birnbaum and Saunders (1969) and Johnson et al. (1995, pp. 651-663). This interest for the BS model is due to its attractive properties, its theoretical arguments, and its relationship with the normal model. Different estimation aspects related to this distribution have been widely studied; see, e.g., Johnson et al. (1995, pp. 656-660). However, no much attention has been payed in new parameterizations of the BS model. We consider an interesting reparameterization proposed by Ferrari and Cribari-Neto (2004) applied to regression models. We propose a regression model for the BS model based on this reparameterization and a methodology similar to that used in linear generalized models. Specifically, let $T_1, \ldots, T_n$ be r.v.'s with $T_i \sim \mathcal{BS}(\mu_i, \delta)$, which p.d.f. is

$$f(t_i;\mu_i,\delta) = \frac{e^{\frac{\delta}{2}}\sqrt{\delta+1}}{4\sqrt{\pi\mu_i}} t_i^{-3/2}\left(t_i + \frac{\delta\mu_i}{\delta+1}\right) e^{-\frac{\delta}{4}\left(\frac{t_i(\delta+1)}{\delta\mu_i}+\frac{\delta\mu_i}{t_i(\delta+1)}\right)}, t_i > 0, \quad (1)$$

where $E(T_i) = \mu_i$, $\mathrm{Var}(T_i) = V(\mu_i)/h(\delta)$, with $V(\mu_i) = \mu_i^2$, and $h(\delta) = (\delta+1)^2/(2\delta+5)$. A Birnbaum-Saunders regression model (BSM) on the

mean can be defined from (1) using the systematic component

$$g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} = \eta_i, \quad i = 1, 2, \ldots, n, \tag{2}$$

where $\mu_i = g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta})$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ is the regression coefficient vector, with $p < n$, and $\mathbf{x_i} = (x_{i1}, \ldots, x_{ip})^\top$ are the values of p explanatory variables. In addition, we assume that $g(\cdot)$ is a known one-to-one continuously differentiable, positive-value function, usually called link function. For example, $g(\mu) = \log(\mu)$ or $g(\mu) = \sqrt{\mu}$.
An iterative process to get the maximum likelihood (ML) estimates of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \delta)^\top$ may be developed by using, for example, the Fisher scoring method, which leads to the system of equations

$$\boldsymbol{\theta}^{(m+1)} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{W}}^{(m)} \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \tilde{\mathbf{W}}^{(m)} \mathbf{z}^{*(\mathbf{m})}, \quad \text{where}$$

$$\tilde{\mathbf{X}} = \begin{pmatrix} \mathbf{X} & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix}, \quad \tilde{\mathbf{W}} = \begin{pmatrix} \mathbf{D(v)} & \mathbf{D(a)s} \\ \mathbf{s}^\top \mathbf{D(a)} & \mathrm{tr}(\mathbf{D(u)}) \end{pmatrix}, \quad \text{and}$$

$$\mathbf{z}^{*(\mathbf{m})} = \tilde{\mathbf{X}} \boldsymbol{\theta}^{(\mathbf{m})} + \{\tilde{\mathbf{W}}^{(\mathbf{m})}\}^{-\mathbf{1}} \begin{pmatrix} \mathbf{D(a)}^{(m)} & \mathbf{0} \\ \mathbf{0} & \mathrm{tr}(\mathbf{D(b)}) \end{pmatrix} \begin{pmatrix} \mathbf{z}^{(m)} \\ 1 \end{pmatrix},$$

with $\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x_p})$, $\mathbf{D(v)} = \mathrm{diag}\{v_1, \ldots, v_n\}$, $\mathbf{D(u)} = \mathrm{diag}\{u_1, \ldots, u_n\}$, $\mathbf{D(a)} = \mathrm{diag}\{a_1, \ldots, a_n\}$, $z = (z_1, \ldots, z_n)^\top$, $\mathbf{s} = (s_1, \ldots, s_n)^\top$,

$$v_i = \frac{\delta}{2\mu_i^2} \frac{1}{\{g'(\mu_i)\}^2} + \frac{\delta^2}{(\delta+1)^2} \frac{1}{\{g'(\mu_i)\}^2} I(\delta),$$

$$u_i = \frac{(\delta^2 + 3\delta + 1)}{2\delta^2(\delta+1)^2} + \frac{\mu_i^2}{(\delta+1)^4} I(\delta), \quad a_i = \frac{1}{g'(\mu_i)},$$

$$z_i = -\frac{1}{2\mu_i} + \frac{\delta}{(\delta t_i + t_i + \delta\mu_i)} + \frac{t_i\,(\delta+1)}{4{\mu_i}^2} - \frac{\delta^2}{4t_i\,(\delta+1)},$$

$$s_i = \frac{1}{2\mu_i(\delta+1)} + \frac{\delta\mu_i}{(\delta+1)^3} I(\delta) \quad \text{and}$$

$$I(\delta) = \int_0^\infty \frac{1}{4} e^{\frac{\delta}{2}} \sqrt{\delta} e^{-\frac{1}{4}\left(\frac{(\delta+1)t}{\delta\,\mu} + \frac{\delta\,\mu}{(\delta+1)t}\right)\delta} \left(t + \frac{\delta\,\mu}{\delta+1}\right)^{-1} \frac{t^{-3/2}}{\sqrt{(\pi\,\delta\,\mu)/(\delta+1)}} dt.$$

2 Local influence

The idea behind local influence is concerned with the study of the behaviour of some influence measure around the non-perturbed vector $\boldsymbol{\omega_0}$. For example, if the likelihood displacement $\mathrm{LD}(\omega) = 2\{\mathrm{L}(\hat{\boldsymbol{\theta}}) - \mathrm{L}(\hat{\boldsymbol{\theta}}_\omega)\}$ is used, where $\hat{\boldsymbol{\theta}}_\omega$ denotes the ML estimate under the perturbed model, the suggestion of Cook (1986) is to investigate the normal curvature of the lifted

line $\mathrm{LD}(\omega_0 + a\boldsymbol{\ell})$, where $a \in I\!R$, around $a = 0$ for an arbitrary direction $\boldsymbol{\ell}$, where $||\boldsymbol{\ell}|| = 1$. He showed that the normal curvature may be expressed in the general form $\mathrm{C}_\ell(\boldsymbol{\theta}) = 2|\boldsymbol{\ell}^\top \boldsymbol{\Delta}^\top \ddot{\mathbf{L}}_{\theta\theta}^{-1} \boldsymbol{\Delta}\boldsymbol{\ell}|$, where $\boldsymbol{\Delta}$ is a $(p+q) \times s$ matrix with elements $\Delta_{ij} = \partial^2 \mathrm{L}(\boldsymbol{\theta}|\boldsymbol{\omega})/\partial\theta_i \partial\omega_j$, for $i = 1, \ldots, p+q$ and $j = 1, \ldots, s$, with all the quantities evaluated at $\hat{\boldsymbol{\theta}}$. Lesaffre and Verbeke (1998) suggested evaluating the normal curvature at the direction of the ith observation, This consists of evaluating $\mathrm{C}_\ell(\boldsymbol{\theta})$ at an $n \times 1$ vector $\boldsymbol{\ell}_i$ formed by zeros with one at the ith position. Suppose the log-likelihood function for $\boldsymbol{\theta}$ is expressed as $\mathrm{L}(\boldsymbol{\theta}|\boldsymbol{\omega}) = \sum_{i=1}^{n} \omega_i \mathrm{L}_i(\mu_i, \delta)$, where $0 \leq \omega_i \leq 1$ is the weight of the ith case. Under this perturbation scheme, the matrix $\boldsymbol{\Delta}^\top$ takes the form $\boldsymbol{\Delta}^\top = [\boldsymbol{\Delta}_{\boldsymbol{\beta}}^\top, \boldsymbol{\Delta}_{\boldsymbol{\delta}}^\top]^\top$, where $\boldsymbol{\Delta}_{\boldsymbol{\beta}}$ is a $p \times n$ matrix expressed as $\boldsymbol{\Delta}_{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{D}(\mathbf{a})\mathbf{D}(\mathbf{e})$, with $\mathbf{D}(\mathbf{e}) = \mathrm{diag}\{e_1, \ldots, e_n\}$ and $e_i = -\frac{1}{2\mu_i} + \left(\frac{\delta+1}{\delta} t_i + \mu_i\right)^{-1} + \frac{(\delta+1)}{4\mu_i^2} t_i - \frac{\delta^2}{4(\delta+1)t_i}$. In addition, $\boldsymbol{\Delta}_{\boldsymbol{\delta}} = (b_1, \ldots, b_n)^\top$, where $b_i = \frac{1}{2} + \frac{1}{2(\delta+1)} + \frac{(t_i+\mu_i)}{(\delta t_i + t_i + \delta\mu_i)} - \frac{t_i}{4\mu_i} - \frac{\delta(\delta+2)\mu_i}{4(\delta+1)^2 t_i}$.

3 Application

To illustrate the obtained results, we consider the biaxial fatigue data set ($n = 46$) analyzed by Rieck and Nedelman (1991). Galea et. al (2004) developed various diagnostic methods for a linear regression model under a logarithmic BirnbaumSaunders distribution for the errors. The response variable is given by the number of cycles to the occurrence of the flaw (T) and the explanatory variable is the work for log-cycle ($log(W)$) expressed by $\mathrm{M}_j/\mathrm{m}^3$. We propose to fit the BSM given by $T_i \sim \mathcal{BS}(\mu_i, \delta)$, and

$$\log(\mu_i) = \beta_0 + \beta_1 log(W_i), \quad i = 1, 2, \ldots, 46, \tag{3}$$

The ML estimates $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \delta)^\top = (\beta_0, \beta_1, \delta)^\top$ and standard errors (in parentheses) are, respectively, $\hat{\boldsymbol{\beta}}_0 = 12.360\,(0.391)$, $\hat{\boldsymbol{\beta}}_1 = -1.671\,(0.109)$ and $\hat{\delta} = 11.816\,(2.464)$. Figure 1 shows the scatterplot and fitted line of the data indicating that the straight line is very adjusted. Figures 2 and 3 show residual plot and normal probability plot with envelope, respectively, which do not present any unusual behavior. Figure 4 shows the index plot of $|d_{\max}|$ under the weighted case perturbation scheme, which does not present observations that are significantly highlighted of the others.

Acknowledgments: This study was supported by CNPq and Facepe, Brazil, and FONDECYT 1080326 and DIPUV 50-2007, Chile.

References

Birnbaum, Z.W., Saunders, S.C. (1969). A new family of life distribution. *Journal of Applied Probability*, **6**, 319-327.

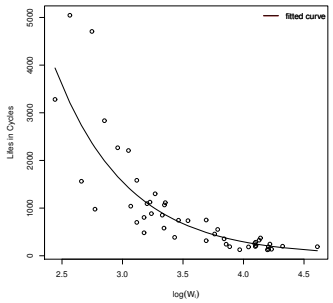


FIGURE 1. Scatterplot for fatigue data and fitted curve.

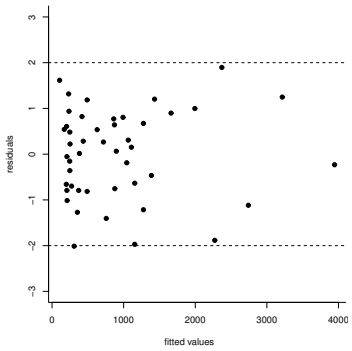


FIGURE 2. r_i^* vs. fitted values plot.

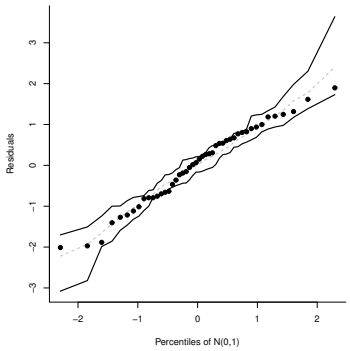


FIGURE 3. Normal probability plot with envelope.

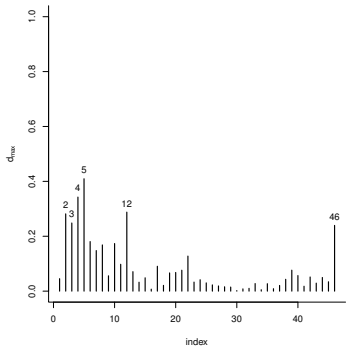


FIGURE 4. index plot of $|d_{\max}|$ under case weight scheme.

Cook, R.D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society, Series B*, **48**, 133-169.

Ferrari, S., Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**, 799-815.

Galea, M., Leiva, V., Paula, G.A. (2004). Influence diagnostics in log-Birnbaum-Saunders regression models. *Journal of Applied Statistics*, **31**, 1049-1064.

Johnson, N., Kotz, S., Balakrishnan, N. (1995). *Continuous Univariate Distributions*. Volume 2. Second edition. John Wiley and Sons, New York.

Lesaffre, F., Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics*, **38**, 963-974.

Rieck, J.R., Nedelman, J.R. (1991). A log-linear model for the Birnbaum-Saunders distribution. *Technometrics*, **33**, 51-60.

Recovering gene-networks using l_1 and l_0 penalties

Johan J. de Rooi[1,2], Paul H. C. Eilers[2]

[1] Department of Bioinformatics, Erasmus Medical Centre, Rotterdam, The Netherlands, email: j.derooi@erasmusmc.nl
[2] Department of Biostatistics, Erasmus Medical Centre, Rotterdam, The Netherlands

Abstract: One of the goals in the analysis of gene-expression data is to recover networks of interacting genes. An appealing way to derive these networks is to shrink the inverse covariance matrix and present the non-zero relations in a graphical model. Shrinkage of the covariance matrix is done with various penalties, however in most cases with the l_1. A known problem with this penalty is that the networks it produces are often still to 'rich'. What we propose here is to apply the l_0 penalty in order to produce stronger shrinkage, resulting in more nodes having only one or few connections. The algorithm is applied on a gene-expression dataset consisting of 300 genes, derived from 79 brain tumor samples. Preliminary results showing that applying the l_0 yields a sparser structure compared to the l_1 penalty.

Keywords: shrinkage; covariance estimation; graphical models, gene-networks

1 Introduction

During the last years, one of the primary goals in the analysis of gene expression data has been to recover networks of interacting genes. The task is to estimate the relations between the genes in the initial matrix of expression values, and translate these into a graphical model. In such a model each gene is represented as a node, a relation between two genes is visualized by an edge between both. Because genetical networks are considered sparse, most of the nodes of the final network should be connected to a single or only a few other nodes These sparse networks are also appealing in terms of parsimony and as a result returns a model that is interpretable. However, translating the data into a sparse graphical model is a non-trivial task. This because of the high dimensionality of the data and the resulting difficulties with estimating the conditional relations between the nodes. Often applied methods to derive a network are Support Vector Machines, Bayesian networks or methods based on information theory. From a statistical point of view it is logical to use the covariance matrix of the genes to build the network. More elegant is to use the inverse of the covariance

matrix (aka the precision matrix), because of its close relation with partial correlations. Although the final model should be sparse the initial covariance matrix doesn't contain any zeros. While due to its large number of variables and often a limited number of samples the inverse cannot be calculated. In order to derive an invertible covariance matrix and reach a sparse model, shrinkage procedures are applied (see e.g. Friedman et al. 2008). The approach we take fits a regression model on one variable in the model with all others being the predictors (see e.g. Meinshausen and Bühlmann, 2006). This procedure is repeated for all variables. From the regression coefficients we can calculate the the partial correlation as

$$\hat{\rho}_{ij} = \text{sign}(\hat{\beta}_{ij})\sqrt{\hat{\beta}_{ij}\hat{\beta}_{ji}}. \tag{1}$$

Because in the setting with $p > n$ (with p being the number of genes and n the number of samples) the sign of $\hat{\beta}_{ij}$ can differ from the sign of $\hat{\beta}_{ji}$, we define that (1) only holds if $\text{sign}(\hat{\beta}_{ij}) = \text{sign}(\hat{\beta}_{ji})$, if not the coefficient is zero (Krämer et al. 2009). In the next section, the penalties are explained, while in the third paragraph an application will be shown.

2 Variable selection through shrinkage

Shrinkage estimation in statistics is often applied in pursuit of a better predictive model. In the context of high dimensional data the same techniques are applied to derive a model that is sparse, i.e. many relations are set to zero. With putting a coefficient to zero we assume conditional independence, in terms of the network this means the absence of an edge between two nodes given the other nodes in the model. Probably the most often used penalty is the l_1 penalty, aka the lasso. This penalty is attractive because it is convex and does both shrinkage and variable selection by setting coefficients to zero. In connection to recovering sparse networks the l_1 often remains to many edges in the network. In order to further reduce the number of nodes with many relations, we use the l_0 penalty (e.g. Eilers, 2009) and in this way yields a model that better resembles the very sparse nature of genetical networks. Because of the non-convexity of the l_0 penalty we adopt a two step strategy. As a first step the l_1 penalty is applied until a satisfactory solution is reached. In the second step the l_0 penalty is put to work on the remaining non-zero coefficients in the model. With only the l_1 penalty included, the model looks as follows:

$$\hat{\beta}_{lasso} = \underset{\beta}{\text{argmin}}\, ||\mathbf{X}^T\mathbf{X}\beta - \mathbf{y}||_2^2 + \lambda||\beta||_1, \tag{2}$$

with λ being the tuning parameter. For an easy and flexible implementation of the l_1 penalty we use an approximation: $|\beta_j| = \beta_j^2/|\beta_j|$. Using $\beta_j \approx$

$\sqrt{\beta_j^2 + c^2}$, with c being a small constant, gives us the following equation:

$$\beta_{new} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{V}^{-1})^{-1}\mathbf{X}^T\mathbf{y}, \tag{3}$$

with $\mathbf{V} = \text{diag}(\sqrt{\beta^2 + c^2})$, iterating gives us the optimal β (see e.g. Osborne et al., 1999). In case of the l_0 penalty $\mathbf{V}$ is replaced with: $\mathbf{W} = \text{diag}(\beta^2 + c^2)$. The final estimates from the iterative l_1 procedure are used as a starting point for optimizing l_0. An important aspect of the model is the optimization of the tuning parameter which is often done via cross-validation or using an information criterion.

3 Application

In this section we apply the algorithm on a dataset composed of the 300 most varying genes from 79 microarray expression samples, which forms a molecular subgroup in a larger study (Gravendeel et al., 2009). Figure 1 shows the cumulative distribution function for two different network densities. The density is expressed by the ratio of total number of nodes against the number of edges left in the network, the cdf itselves gives away information of how these edges are organised. The figure summarizes how the nodes are connected in the different models, given a common density of the network. In both cases one model is fit with the l_1 penalty and one with the l_0 penalty. In the left panel the ratio of nodes to edges is 1:2.15, in the right panel fewer edges are left and give a ratio of 1:1.25. A steep curve means that many nodes have none or only a few connections. In both panels we see that applying the l_0 penalty yields a network that has more nodes with only a few connections. In other words: applying the l_0 means that the edges are more spread out over the nodes in the model.

4 Discussion

The aim of the procedures as presented here and in the literature are not to recover one 'true' network underlying the data, but rather to unveil the strongest conditional relations present in the data. Preliminary analyses on both real data and small simulations return promising results with respect to the l_0 penalty. However, for a better assessment of the performance of the algorithm proposed more thorough simulations are needed. Next to the fixed lambda as applied here, the application of a 'progressive lambda', dependent on the initial coefficients is investigated.

References

Eilers, P.H.C. (2009). Deconvolution of spike trains using an l_0 penalty. In: *Proceedings of the 24th International Workshop on Statistical Modelling*. 130-137, Ithaca, New York.

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

Gravendeel, L. A. M., Kouwenhoven, M. C. M., Gevaert O. et al. (2009). Intrinsic Gene Expression Profiles of Gliomas Are a Better Predictor of Survival than Histology. *Cancer Research* **69(23)**, 9065-9072.

Krämer, N., Schäfer, J., and Boulesteix, A. (2009). Regularized estimation of large-sclae gene association networks using graphical Gaussian models. *BMC Bioinformatics*, **10**:384.

Meinshausen, N., Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, **34**, 1436-1462.

Osborne, M.R., Presnell, B., and Turlach, B. A. (1999). On the LASSO and its dual. *Journal of Computational & Graphical Statistics*, **9(2)**, 319-337.

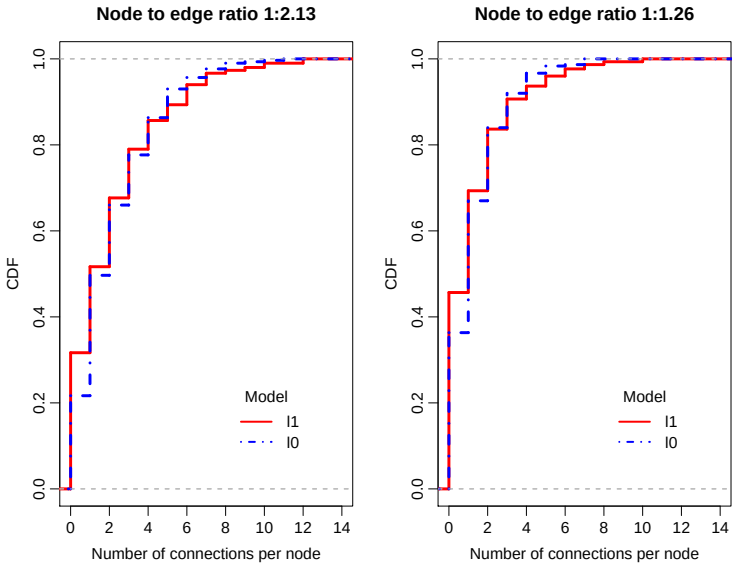


FIGURE 1. Two cdf's for the 300 node matrix.

Estimating the prevalence and the force of infection of Parvovirus B19 in Belgium using hierarchical Bayesian mixture models

Emanuele Del Fava[1], Ziv Shkedy[1]

[1] Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Agoralaan, 3590 Diepenbeek, Belgium.
E-mail: emanuele.delfava@uhasselt.be

Abstract: The prevalence and the force of infection of infectious diseases are usually estimated from cross-sectional prevalence data, obtained by means of some standard cut-off points applied to the antibody levels. However, this cut-off can be chosen arbitrarily or based on simplistic assumptions. We propose to estimate the two epidemiological parameters using a hierarchical Bayesian mixture model with two Gaussian components for the antibody levels, where the cut-off is not chosen *a priori*, but it is given directly by the data. The prevalence model is applied to a cross-sectional sample containing information about antibody levels for Parvovirus B19 for 3094 subjects in Belgium in 2002.

Keywords: Bayesian analysis, mixture models, prevalence, force of infection.

1 Introduction

In the epidemiology of infectious diseases, two important parameters are the prevalence π and the force of infection λ (FOI). The former is the proportion of seropositive individuals in a given population, whereas the latter is the individual risk for a susceptible to naturally acquire an infection. Assuming time homogeneity due to pre-vaccination equilibrium, the two parameters depend on the age of the host:

$$\pi(a) = 1 - \exp\left\{-\int_0^a \lambda(s)ds\right\} \qquad \lambda(a) = \frac{1}{1-\pi(a)}\frac{\partial \pi(a)}{\partial a}. \qquad (1)$$

The prevalence and the force of infection can be estimated using cross-sectional serological data: we obtain the prevalence from the serological binary data and then we derive the force of infection using parametric or nonparametric models (Farrington, 1990; Shkedy *et al.*, 2003; Shkedy *et al.*, 2006).
The main issue in this procedure lies in the choice of the cut-off method. Usually, it is a fixed cut-off given by the producer of the assay (e.g., ELISA),

which divides a single population between susceptible and infected. However, it is more reasonable to think that the observations come from two distinct subpopulations with different parameters, one group for the susceptible individuals and the other for the infected individuals, rather than originating from the same population. Moreover, the fixed cut-off from standard assays has the specificity larger than sensitivity, because it is mainly used for diagnoses, with focus on true negative cases; instead, for serological purposes, a higher sensitivity is required, because the focus is on the prevalence of true positive cases (Vyse *et al.*, 2004).
In this paper we propose a hierarchical Bayesian mixture model (Diebolt and Robert, 1994) for the antibody levels in order to classify the individuals between susceptible and infected and to estimate the prevalence and the force of infection at the same time. The advantage is that we do not have to choose any cut-off point, but it is rather the mixture that classifies the observations by assigning a probability to each observation to belong to a specific component.

2 Methods

The ith subject in the sample is classified either as susceptible or as infected according to a Gaussian mixture model with two components. The distribution of the antibody levels Y_i, $i = 1, \ldots, n$, is given by:

$$Y_i \sim (1 - \pi(a_i))N(\mu_1, \sigma_1^2) + \pi(a_i)N(\mu_2, \sigma_2^2). \tag{2}$$

Here, μ_1 and σ_1^2 are the mean and variance of the susceptible component, while μ_2 and σ_2^2 are the parameters for the infected component; $\pi(a_i)$ is the mixture probability assumed to be age-dependent. We introduce a latent classification variable, Z_i, which is Bernoulli distributed and can be interpreted as the current status of the ith subject:

$$Z_i \sim \begin{cases} 1 & \pi(a_i) \quad \text{infected,} \\ 0 & 1 - \pi(a_i) \quad \text{susceptible.} \end{cases} \tag{3}$$

We interpret the mixing probability $\pi(a_i)$ as the weight of the "infected" component at age a_i and corresponds to the prevalence, while the probability $1 - \pi(a_i)$ is the weight of the "susceptible" component at age a_i (Vyse *et al.*, 2004). We model the probability $\pi(a)$ using three non decreasing parametric models: a Weibull model for prevalence with constant FOI, a log-logistic model for prevalence and FOI, and a nonlinear model, introduced by Farrington (1990), for prevalence and FOI. Here follow the three models:

$$\pi(a) = 1 - \exp(-\gamma a) \qquad \lambda(a) = \gamma. \tag{4}$$

$$\pi(a) = \frac{e^{\gamma_0} a^{\gamma_1}}{1 + e^{\gamma_0} a^{\gamma_1}} \qquad \lambda(a) = \frac{e^{\gamma_0} \gamma_1 a^{\gamma_1 - 1}}{1 + e^{\gamma_0} a^{\gamma_1}}. \tag{5}$$

$$\begin{aligned} \pi(a) &= 1 - \exp\left\{ \frac{\beta_1}{\beta_2} a e^{\beta_2 a} + \frac{1}{\beta_2} \left(\frac{\beta_1}{\beta_2} - \beta_3 \right) \left(e^{-\beta_2 a} - 1 \right) - \beta_3 \right\} \\ \lambda(a) &= (\beta_1 a - \beta_3)\, e^{-\beta_2 a} + \beta_3. \end{aligned} \tag{6}$$

All parameters of interest, for mixture, prevalence and FOI, are estimated within the hierarchical Bayesian setting. We use diffuse but proper prior distributions to express our uncertainty about the true value of the parameters. We constrain the mean level μ_1 of susceptible to be smaller than the mean level μ_2 of infected :

$$\mu_j \sim N(0, 10^{-3}) \quad j = 1, 2 \quad \text{with } \mu_1 < \mu_2. \tag{7}$$

The diffuse prior for the variances σ_j^2 is an Inverse Gamma, but for computational reasons we express it as a diffuse Gamma for the precision $\tau_j = 1/\sigma_j^2$:

$$\tau_j \sim \Gamma(10^{-2}, 10^{-2}) \quad j = 1, 2. \tag{8}$$

Finally, the coefficients of the parametric models for the prevalence have a diffuse normal prior distribution. Namely, for the the log-logistic models we have:

$$\gamma_j \sim N(0, 10^{-3}) \quad j = 1, 2. \tag{9}$$

3 Results

The mixtures are fitted through MCMC simulation in JAGS using 3 chains of 20000 iterations, whose first 4000 iterations were discarded as burn-in part. The estimates of μ_j and σ_j^2 under the three models are very close to each other (see Table 1). Figure 1 shows the posterior mean of the prevalence and the force on infection obtained from the log-logistic model, together with the proportions seropositive estimated from the posterior median of the current status variable Z_i. We see that the prevalence increases monotonically with the age and the force of infection has a peak at the beginning (around 6 months) and then declines.

4 Sensitivity analysis

We performed a sensitivity analysis by using different distributions for the data, namely, a Gamma and a Student t distribution. Indeed, in principle, a

TABLE 1. Posterior means of the μ_j and σ_j^2 of the normal mixture under the three parametric models.

Par.	Weibull const. FOI	Log-logistic	Farrington
μ_1	5.07 (5.04,5.10)	5.07 (5.04,5.10)	5.07 (5.04,5.08)
μ_2	1.65 (1.63,1.68)	1.66 (1.63,1.68)	1.66 (1.63,1.68)
σ_1	0.63 (0.61,0.66)	0.63 (0.61,0.65)	0.63 (0.61,0.65)
σ_2	0.39 (0.37,0.41)	0.39 (0.37,0.41)	0.39 (0.37,0.41)

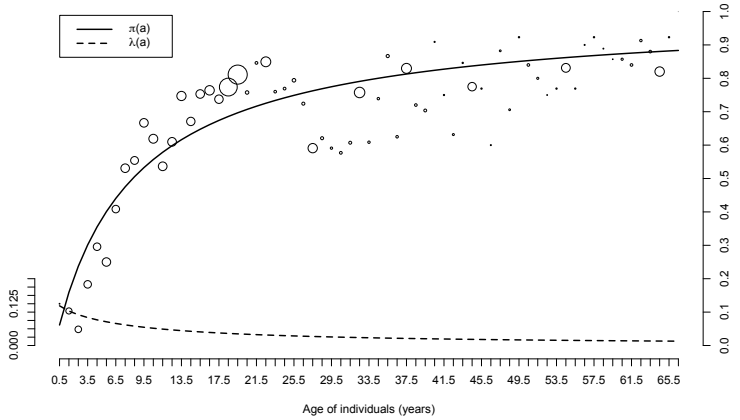


FIGURE 1. Posterior mean of the prevalence and force of infection according to the log-logistic model and seropositive proportions from the mixture model.

Gamma distribution can take into account the skewness of the data, while a Student t with low degrees of freedom is more robust than the normal distribution. In Figure 2 we show the three estimated mixture models with mixing weights equal to the average prevalence, obtained using the log-logistic model. We notice, on the one hand, that all the distributions can fit quite well the seronegative component, while, on the other hand, none of them is able to capture the strong skewness of the seropositive group. The Gaussian and the Gamma distributions almost coincide, whereas the Student t is more shifted towards right, with higher means and smaller variances (see Table 2); the Student t is characterized by low degrees of freedom, i.e., $\nu = 5.60(4.37, 7.31)$ and this implies that the distribution is more robust than the Gaussian and can fit better the thick tails of the two components.

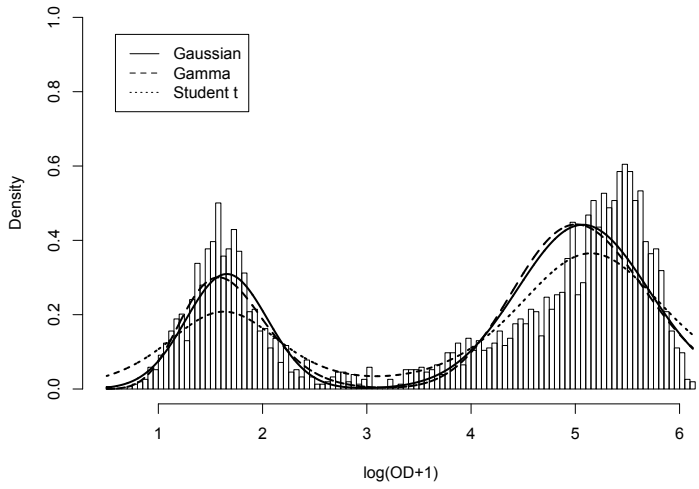


FIGURE 2. Histogram of the data with over imposed the estimated mixture models using a Gaussian distribution (solid line), a Gamma (dashed line), and a Student t (dotted line).

TABLE 2. Comparison of the posterior mean and variance of the two components under the Gaussian, the Gamma, and the Student t distributions.

Par.	Gaussian	Gamma	Student t
μ_1	5.07 (5.04,5.10)	5.09 (5.06,5.12)	5.15 (5.12,5.18)
μ_2	1.66 (1.63,1.68)	1.68 (1.65,1.71)	1.63 (1.60,1.65)
σ_1	0.63 (0.61,0.65)	0.63 (0.62,0.65)	0.53 (0.50,0.56)
σ_2	0.39 (0.37,0.41)	0.41 (0.40,0.43)	0.31 (0.29,0.34)

5 Discussion and future research

In conclusion, we found that mixture models can be useful when a cut-off point is unknown or chosen arbitrary and the advantage of the Bayesian approach is to provide the posterior distribution of each parameter. Mixture models do not always provide better results than a fixed cut-off, namely, in case the components are mostly overlapping, but the comparison of scenarios can give a deeper insight in the data. Some issues are related to the correct choice of the distribution of the data and to the absence of a selection criterion for the best model (Celeux *et al.*, 2006). A possible ex-

planation for the strong skewness of the seropositive component could be that the mean antibody level of seropositive subjects is not constant, but it rather decreases with age, perhaps due to waning in the antibody levels of older individuals. Therefore, a mixture model that took into account this phenomenon could provide a better fit to the data. Other developments are feasible, for instance, using a constrained unstructured mean structure for the prevalence and FOI in order to have more flexibility or extending the methodology to a multivariate setting.

References

Diebolt, J., and Robert, C.P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. R. Statist. Soc. B*, **56**, 363-375.

Farrington, C.P. (1990). Modeling forces of infection for measles, mumps and rubella. *Statistics in Medicine*, **9**, 953-967.

Shkedy, Z., Aerts, M., Molenberghs, G., Beutels, Ph., and Van Damme, P. (2003). Modelling forces of infection by using monotone local polynomials. *Applied Statistics*, **52**, 469-485.

Shkedy, Z., Aerts, M., Molenberghs, G., Beutels, Ph., and Van Damme, P. (2006). Modelling age-dependent force of infection from prevalence data using fractional polynomials. *Statistics in Medicine*, **25**, 1577-1591.

Vyse, A.J., Gay, N.J., Hesketh, L.M., Morgan-Capner, P., and Miller, E. (2004). Seroprevalence of antibody to varicella zoster virus in England and Wales in children and young adults. *Epidemiology and infection*, **132**, 1129-1134.

Celeux, G., Forbes, F., Robert, C.P., and Titterington, D.M. (2006). Deviance information criteria for missing data models (with discussion). *Bayesian Analysis*, **1**, 651-674.

Modeling the Non-Monotonic Association between 25-Hydroxyvitamin D and Mortality in a Representative US Population Sample

R. A. Durazo-Arvizu[1], C. Sempos[2], A. Luke[1], E. A. Yetley[2], B. Dawson-Hughes[3], H. Kramer[1], G. Cao[1], J. T. Dwyer[2], R. L. Bailey[2], A. J. Rovner[4], M. F. Picciano[2]

[1] Department of Preventive Medicine end Epidemiology, Loyola University Chicago, Maywood, IL 60153
[2] Office of Dietary Supplements. National Institutes of Health. Bethesda, IL 20156
[3] Bone Metabolism Laboratory, Tufts University, Boston, MA 02111
[4] Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20816

Abstract: Multivariate logistic regression was applied to model the non-linear, asymmetrical relationship between all cause mortality and 25[OH]D in a national representative sample of the US. Precisely, two analytic strategies were implemented, namely 1) transformation of 25[OH]D to normality and 2) restricted cubic splines. Estimators and standard errors of the nadir of the curve were derived. For all participants, the estimated 25[OH]D of minimum mortality (95% confidence interval) is 99 nmol/L(52-146), and 82 nmol/L (52-111) for the normal transformation and cubic splines approach respectively. The corresponding values for NH-Whites are 91 nmol/L (57-124) and 82 nmol/L (48-116).

Keywords: Vitamin D; Minimum Mortality; 25[OH]D.

1 Introduction

Large population studies and clinical research have implicated vitamin D as a potential risk factor for several chronic diseases. Moreover, studies have demonstrated a non-monotonic association between all-cause mortality and vitamin D, as measured by circulating levels of 25-Hydroxyvitamin D (25[OH]D), with excess mortality at both low and high levels (Melamed et al, 2008). However, the shape of this association appears asymmetric, with a clear upturn on the left and a shallow increase with large 25[OH]D levels(see Figure 1). To our knowledge no attempt has been made to quantitatively estimate the 25[OH]D level at which minimum mortality occurs, accounting for the observed asymmetry. We modeled the non-monotonic, asymmetric association between 25[OH]D and all-cause mortality in the

Third National Health and Nutrition Examination Survey (NHANES III) follow-up study.

The baseline examination for NHANES III took place in the years 1988-1994. Vital status through December 31, 2000 was assessed based on a probabilistic match between personal identifiers from NHANES III and the death certificate records from the National Death Index.

The analytic sample included a total of 17,705 participants ages 17 and older who received the examination including blood draw. Excluded from that sample were those missing information on vital status (n=11), women who were pregnant at baseline (n=338), less than 20 years of age (n=1,082), and missing data for serum 25[OH]D (n=765), serum creatinine (n=344), body mass index (n=34) and systolic blood pressure (n=25) for a total analytic sample size of 15,106 participants. The sample included 7,221 men and 7,885 women aged 20 and older. There were a total of 2,264 deaths during the follow-up period with 1,271 deaths among men and 993 among women.

Results are reported for all participants, and for Non-Hispanic Whites. Logistic regression adjusting for age, education level, glomerular filtration rate (GFR), season, body mass index (BMI), sex, race/ethnicity, medication use (anticonvulsants, estrogens, loop diuretics and/or thiazide diuretics), smoking habits, and systolic blood pressure level was fit to the data. Two statistical modeling strategies were implemented to accommodate the non-linear, asymmetric relationship between mortality and 25[OH]D, namely 1) transformation of 25[OH]D to normality and 2) restricted cubic splines.

Table 1 displays means, proportions and corresponding standard error estimates of the variables used in the multivariate analyses for all subjects and for Non-Hispanic Whites aged 20 and older.

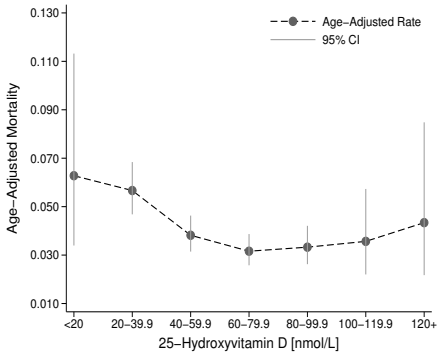


FIGURE 1. Age-adjusted mortality rates by level of 25-Hydroxyvitamin D, along with 95% confidence intervals. NHANES III follow-up study.

TABLE 1. Unadjusted or crude weighted percentages and means with corresponding standard errors. All subjects and Non-Hispanic Whites aged 20 and older.

Variables	**All Subjects** $N^{\dagger} = 15,106; D^{\ddagger} = 2,264$	**NH-Whites** N=6,355; D=1,332
% Men	48.5(0.44)	48.7(0.53)
% Winter Season	38.1(3.75)	34.1(3.92)
% Current Smokers	28.4(0.80)	28.6(1.01)
% Medicine Use	15.5(0.48)	17.2(0.60)
Education:		
% Less HS[§]	24.7(1.03)	20.4(1.15)
% Completed HS	33.5(0.73)	34.5(0.90)
Age(yr)	45.1(0.47)	46.4(0.59)
Serum 25[OH]D(nmol/L)	64.5(0.73)	68.9(0.79)
GFR[¶](mL/(min×1.73×m^2))	92.2(0.52)	89.2(0.54)
BMI[*](kg/m^2)	26.6(0.10)	26.4(0.13)
SBP[♯](mmHg)	123(0.40)	123(0.51)

[†]Number of subjects; [‡]Deaths; [§]High School; [¶]Glomerular filtration rate; [*]Body mass index; [♯]Systolic blood pressure.

2 Transformation to Normality

Let $X = 25[OH]D$ and X^2 be independent variables in a logistic regression model fit to the data to capture the 25[OH]D-Mortality association, which adjusts for confounders $\mathbf{Z}$. Precisely,

$$Pr(Death \mid X, \mathbf{Z}) = \frac{1}{1 + \exp\left[\alpha_0 + \alpha_1 X + \alpha_2 X^2 + \mathbf{\Omega}'\mathbf{Z}\right]}$$

where $\mathbf{\Omega}$ is a vector of unknown parameters to be estimated, associated with the vector of confounders $\mathbf{Z}$. This model presumes that the risk of death, as a function of 25[OH]D, is symmetrical. This amounts to assuming that small and large values of 25[OH]D have similar effects on all-cause mortality. Figure 1 appears to contradict such an assumption. The curve to the right of the nadir is shallower than to the left of it, suggesting an asymmetrical association. Fitting a symmetrical function to an asymmetrical relationship can lead to inconsistent findings as demonstrated previously (Goetghebeur et al, 1995).

We propose an approach based on a transformation of the independent variable. Cornfield, Gordon and Smith (Cornfield et al, 1961), and later Kay and Little (Kay et al, 1987) argue, in the univariate case, that the best transformation can be motivated by noticing that $Pr(Death \mid Y) =$

$\frac{1}{1+\frac{q}{p}\frac{f_0(y)}{f_1(y)}}$, with $p = Pr(Death)$, $q = 1-p$, and the distribution of Y among survivors and not survivors denoted by f_0 and f_1 respectively. Thus, the logistic regression model will be a good fit of the data when $f_0(y)/f_1(y)$ is a function of the form $\exp(\psi(y))$, where ψ is an arbitrary function of y. In particular, if the distribution of Y for survivors and non-survivors follows a normal distribution, then

$$\frac{f_0(y)}{f_1(y)} = C \exp\left\{ y\left[\frac{\mu_0}{\sigma_0^2} - \frac{\mu_1}{\sigma_1^2}\right] + y^2\left[\frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_0^2}\right]\right\},$$

where C is a constant, $f_0 = N(\mu_0, \sigma_0^2)$, $f_1 = N(\mu_1, \sigma_1^2)$. As a consequence, if the independent variable follows a normal distribution with differing means and standard deviations for survivors and non-survivors, then the logistic regression model including both Y and its square will fit the data well.
We searched for the best transformation of 25[OH]D using the Box-Cox transformation approach and the ladder of powers method suggested by Tukey (Tukey 1957), which resulted in the square root transformation. The best fitting model is given by:

$$Pr(Death \mid X, \mathbf{Z}) = \frac{1}{1 + \exp\left[\alpha_0 + \alpha_1\sqrt{X} + \alpha_2 X + \mathbf{\Omega}'\mathbf{Z}\right]}$$

The parameters in this model are estimated by maximum likelihood methods and the nadir of the curve is computed as a function of the estimated parameters, as follows:

$$25[OH]D_{min} = \text{Nadir}(\hat{\alpha}_1, \hat{\alpha}_2) = \frac{1}{4}\frac{\hat{\alpha}_1^2}{\hat{\alpha}_2^2}$$

Hence, from maximum likelihood theory and applying the delta method

$$\left[\frac{1}{4}\frac{\hat{\alpha}_1^2}{\hat{\alpha}_2^2} - \frac{1}{4}\frac{\alpha_1^2}{\alpha_2^2}\right] \xrightarrow{d} N(0, \sigma_*^2) \quad \text{with} \quad \sigma_*^2 \doteq \nabla \cdot \Sigma_* \cdot \nabla^T$$

$$\text{where } \nabla = \left[\frac{\partial \text{Nadir}(\alpha_1, \alpha_2)}{\partial \alpha_1}, \frac{\partial \text{Nadir}(\alpha_1, \alpha_2)}{\partial \alpha_2}\right] \text{ and } \quad \Sigma_* = Var(\hat{\alpha}_1, \hat{\alpha}_2)$$

3 Restricted Cubic Splines

Let $X_1 = 25[OH]D$, for j=1,2,3

$$X_{j+1} = (X_1 - t_j)_+^3 - \frac{(t_5 - t_j)}{(t_5 - t_4)}(X_1 - t_4)_+^3 + \frac{(t_4 - t_j)}{(t_5 - t_4)}(X_1 - t_5)_+^3$$

where t_1, t_2, t_3, t_4, t_5 are referred as knots, and $(\mathbf{a})_+$ takes the value $\mathbf{a}$ if $\mathbf{a} > \mathbf{0}$ and zero otherwise. Then, the logistic regression relating the probability of death to 25[OH]D adjusted for potential confounders, $\mathbf{Z}$, takes

the following form:

$$\text{logit}\,(25[OH]D) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \mathbf{\Omega'Z}$$

It is shown that if we define β_5, and β_6 by

$$\begin{aligned}\beta_5 &= \frac{1}{(t_5 - t_4)}\left[\beta_2(t_1 - t_5) + \beta_3(t_2 - t_5) + \beta_4(t_3 - t_5)\right] \\ \beta_6 &= \frac{1}{(t_4 - t_5)}\left[\beta_2(t_1 - t_4) + \beta_3(t_2 - t_4) + \beta_4(t_3 - t_4)\right]\end{aligned}$$

then,

$$\frac{d\ \text{logit}}{dX_1} = A_j X_1^2 + B_j X_1 + C_j, \quad X_1 \in (t_j, t_{j+1}], j = 0, ...5 \tag{1}$$

$$A_j = 3\sum_{i=1}^{i=j} \beta_{i+1};\ B_j = -6\sum_{i=1}^{i=j} t_i\beta_{i+1};\ C_j = \beta_1 + 3\sum_{i=1}^{i=j} t_i^2\beta_{i+1}$$

The nadir of the mortality-25[OH]D curve is estimated by $X(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$, where X is one of the solutions to equation 1 satisfying $X \in (t_j, t_{j+1}]$. The standard error estimate of this statistic can be estimated by the delta method. That is,

$$\sqrt{n}\left[X(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4) - X(\beta_1, \beta_2, \beta_3, \beta_4)\right] \xrightarrow{d} N\left(0, \nabla X \cdot \Sigma \cdot \nabla X^T\right)$$

∇X represents the vector of partial derivatives of $X(\beta_1, \beta_2, \beta_3, \beta_4)$ and Σ the variance-covariance matrix of the vector of parameter estimators $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$.

4 Results

For all participants, the estimated 25[OH]D of minimum mortality (95% confidence interval) is 99 nmol/L (52-146), and 82 nmol/L (52-111) for the normal transformation and cubic splines approach respectively. The corresponding values for NH-Whites are 91 nmol/L (57-124) and 82 nmol/L (48-116). We refer to Table 2 for a summary of our main results.

5 Conclusion

The concentration of serum 25(OH)D estimated to be at the minimum risk of death or nadir of the curve was lower using the cubic splines approach compared to the normal transformation approach, corresponding to Figure 1. Furthermore, the cubic splines-based approach yielded narrower confidence intervals. Other modeling strategies will be implemented in future analyses, including fractional polynomials (Royston et al, 1994) and piecewise polynomials (Goetghebeur et al, 1995).

TABLE 2. Estimates of 25[OH]D at the Nadir of the U-shaped curve with risk of death from all causes.

Non-Hispanic Whites, Both Sexes		
Statistical Approach	**Nadir**	**95% CI**
Transformation to Normality	92	57-127
Cubic Splines	83	50-116

All Subjects, Both Sexes		
Statistical Approach	**Nadir**	**95% CI**
Transformation to Normality	100	52-148
Cubic Splines	83	51-115

Acknowledgments: Supported by a grant from the National Institute on Aging, National Institutes of Health (NIH), Bethesda, MD, grant number AG10353, and by an Office of Dietary Supplements administrative supplement to NIH grant number 5R37 HL045508-17.

References

Cornfield J., Gordon T., and Smith W.W. (1961). Quantal response curves for experimentally uncontrolled variables. *Bull Int Stat Inst*, **38**, 97-115

Goetghebeur E.J.T., and Pocock S.J. (1995). Detection and Estimation of J-Shaped Risk-Response Relationships. *Journal of the Royal Statistical Society A*, **158**, 107-121

Kay R., and Little S. (1987). Transformations of the explanatory variables in the logistic regression model of binary data. *Biometrika*, **74**, 495-501.

Melamed M.L., Michos E.D., Post W., and Astor B. (2008). 25-Hydroxyvitamin D and the risk of mortality in the general population. *Arch Intern Med*,**168**,1629-1637

Royston P., and Altman D.G. (1994). Regression Using Fractional Polynomials of Continuous Covariates. Parsimonious Parametric Modelling. *Applied Statistics*, **43**, 429-467.

Tukey J.W. (1957). On the Comparative Anatomy of Transformations. *Annals of Mathematical Statistics*, **28**, 602-632.

Bayesian Modelling of Underreported Count Data

Michaela Dvorzak[1], Gerhard Neubauer[1], Helga Wagner[2]

[1] Institute of Applied Statistics, Joanneum Research, Graz, Austria
(Corresponding author: michaela.dvorzak@joanneum.at)
[2] Department of Applied Statistics, Johannes Kepler University, Linz, Austria

Abstract: Taking a binomial approach for underreported counts results in a number of models that allow to estimate the total number of events under different variability assumptions. In this paper we propose Bayesian estimation for these models and investigate one of them - the beta-Poisson model - in detail. Estimation of the model parameters is accomplished by two MCMC schemes making use of improved auxiliary mixture sampling and Metropolis-Hastings steps. Both methods are compared with respect to their efficiency. The performance of both MCMC methods is illustrated by simulations and applications to real data.

Keywords: Underreporting, MCMC, auxiliary mixture sampling, beta-Poisson model.

1 Introduction

The binomial approach to underreporting allows to decompose the expectation of an observed count as $\mu = \lambda\pi$, where λ is the total number of events and π is the reporting probability. Hence for $\pi < 1$ we do not expect to observe each of the λ events, but only $\mu < \lambda$. As real data often show more variability than the binomial model can handle, mixture models have been proposed, where either one or both binomial parameters are assumed to be random. The resulting marginal models are the beta-binomial, negative binomial or beta-Poisson model. For details see Neubauer, Djuraš and Friedl (2010).
Here we focus on the beta-Poisson model that arises from assuming $Y|L, P \sim \text{Binomial}(L, P)$ together with $L \sim \text{Poisson}(\lambda)$ and $P \sim \text{Beta}(\gamma, \delta)$. The marginal beta-Poisson distribution has the moments $\text{E}(Y) = \mu = \lambda\frac{\gamma}{\gamma+\delta} = \lambda\pi$ and

$$\text{var}(Y) = \mu\left(1 + \frac{\lambda(1-\pi)}{1+\gamma+\delta}\right) = \mu\phi \quad \text{with } \phi \geq 1.$$

While the mean decomposition is preserved, the model obviously allows for larger variability in the data as $\text{var}(Y) > \mu$.

The marginal probability mass function of the beta-Poisson model is given by

$$f(y|\lambda,\gamma,\delta) = \frac{\lambda^y}{y!}\frac{B(y+\gamma,\delta)}{B(\gamma,\delta)}\ {}_1F_1\left[y+\gamma, y+\gamma+\delta, -\lambda\right]$$

where ${}_1F_1$ denotes the confluent hypergeometric function.
To provide the model with greater flexibility we allow for semi-parametric regression by using $E(Y_t) = \lambda_t\pi$, $t = 1,\ldots,T$, where $\log(\lambda_t) = \tilde{x}_t'\tilde{\beta} + f(t)$. Here $\tilde{x}_t$ is a d-dimensional vector of regressors and $\tilde{\beta}$ the corresponding vector of unknown parameters. $f(t)$ is a smooth function which we specify using P-splines (Eilers and Marx, 1996) as $f(t) = u_t'\alpha$, where u_t is the vector of m B-spline basis functions evaluated at t and α is the coefficient vector. Thus, $\log(\lambda_t)$ has the linear structure

$$\log(\lambda_t) \ = \ \tilde{x}_t'\tilde{\beta} \ + \ u_t'\alpha \ = \ x_t'\beta,$$

where $x_t = (\tilde{x}_t, u_t)$ and $\beta = (\tilde{\beta}, \alpha)$.

2 Bayesian inference for underreporting models

Bayesian estimation of underreporting can be carried out by MCMC methods. One strategy to estimate the underreporting models discussed above is to make use of improved auxiliary mixture sampling for binomial, Poisson and negative binomial data as discussed in Frühwirth-Schnatter et al. (2009) in combination with Metropolis-Hastings (MH) steps. An alternative for estimation of these models is to use a pure MH algorithm. Both methods will be applied in this paper to estimate the beta-Poisson model.

2.1 Bayesian inference for the beta-Poisson model

As the posterior of the parameter vector $\vartheta = (\beta,\gamma,\delta)$ in the beta-Poisson regression model cannot be derived analytically, estimation of the model parameters is accomplished using a MCMC scheme. We consider two such schemes, where the first uses auxiliary mixture sampling together with MH steps, and the second is a pure MH algorithm.

Prior settings

For both methods the same prior settings are used. To express vague prior knowledge we use a proper but uninformative normal prior for β with zero mean and covariance matrix Σ that reduces to cI ($c = 100$) for a strictly parametric model. We use a second order random walk prior for the spline coefficients α and specify the prior for the smoothing parameter as Inverse Gamma$(0.1, 0.1)$. For $\pi = \gamma/(\gamma+\delta)$ and $\theta = \gamma+\delta$ we use the priors Beta$(1,1)$ and Gamma$(1.5,1)$, respectively. These priors are relatively uninformative but specific prior knowledge could be incorporated by using different parameters.

Method 1: Improved Auxiliary Mixture Sampling

Considering the conditional Poisson regression model $Y_t|P_t \sim \text{Poisson}(\lambda_t P_t)$ we can use the improved auxiliary mixture sampler for count data to sample the vector of regression coefficients β (Frühwirth-Schnatter et al, 2009). If the predictor contains a P-spline part, then the associated smoothing parameter is sampled from an inverse Gamma distribution. The parameters π and θ are sampled by two MH steps. We use a uniform random walk MH for π, and a log-normal random walk MH for θ. In a further step the reporting probabilities P_t are sampled from a beta distribution.

Method 2: Metropolis-Hastings algorithm

In our second sampling scheme the auxiliary mixture sampling step of method 1 is replaced by another MH step to sample the vector β. For strictly parametric models we use a proposal density which is motivated by considering the beta-Poisson model as a quasi-Poisson model with $E(Y_t) = \exp(\eta_t)$ and $\eta_t = x_t'\rho = \rho_0 + x_{t1}\rho_1 + \ldots + x_{tk}\rho_k$. In our underreporting model $\eta_t = \log(\lambda_t \pi)$ and therefore, the linear predictor is given as

$$x_t'\rho = \log \pi + x_t'\beta$$

with $\rho_j = \beta_j$, $j = 1, \ldots, k$ and $\rho_0 = \beta_0 + \log \pi$. A proposal for β is obtained by fitting a quasi-Poisson model to the data where the resulting estimate $\hat{\rho} = (\hat{\rho}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ asymptotically follows a normal distribution with mean ρ and covariance matrix Σ. Conditional on π, the proposal for β is multivariate normal with mean $(\hat{\rho}_0 - \log \pi, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ and scaled covariance matrix $\hat{\phi}\hat{\Sigma}$ obtained from the fitted quasi-Poisson model. For semi-parametric regression the method proposed in Brezger and Lang (2006) can be used. Sampling of π and θ is carried out as in method 1.

3 Application to simulated data

The performance of both MCMC methods is investigated in a simulation study where the degree of Poisson overdispersion is varied. We use simulated data from a beta-Poisson distribution using $\lambda_t = \exp(\beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 \sin(\pi t/\psi))$ for $t = 2, 4, 6, \ldots, 200$. $\beta = (4, 0.01, -0.00005, 0.1)$ denotes the parameter settings we use and $\psi = 365.25/7$ tunes the trigonometric function. To obtain different data situations we use combinations of $\pi = (0.5, 0.7, 0.9)$ and $\gamma = (1/2, 1, 4, 8)$, resulting in 12 simulation settings. For each setting 100 samples are drawn, and the samplers are run for 10000 iterations (after a burn-in of 10000) for each data set.
The left part of Table 1 shows the simulation settings, i.e. the values of the beta parameters (γ, δ), the first two moments of the beta variable and the resulting Poisson overdispersion parameter ϕ.

TABLE 1. Effective sample size

	Simulation settings				Method 1			Method 2		
	(γ, δ)	E(P)	var(P)	ϕ	π	β_0	θ	π	β_0	θ
a	$(1/2, 1/2)$	0.50	0.13	20.34	369	318	431	391	283	358
b	$(1, 1)$		0.08	13.89	215	232	251	326	490	337
c	$(4, 4)$		0.03	5.30	91	132	131	126	187	168
d	$(8, 8)$		0.02	3.28	98	151	189	133	204	236
e	$(1/2, 3/14)$	0.70	0.12	14.54	580	517	740	502	215	440
f	$(1, 3/7)$		0.09	10.56	390	395	494	442	459	446
g	$(4, 12/7)$		0.03	4.46	146	251	256	226	408	334
h	$(8, 24/7)$		0.02	2.87	107	170	261	138	217	291
i	$(1/2, 1/18)$	0.90	0.06	5.97	723	724	935	460	257	448
j	$(1, 1/9)$		0.04	4.66	489	618	725	399	400	497
k	$(4, 4/9)$		0.02	2.42	189	386	513	247	508	559
l	$(8, 8/9)$		0.01	1.78	182	408	728	214	473	760

Both methods yield reasonable estimates for π, θ and β_0 in most settings. For settings with large θ we approach the Poisson limit, where the model is not identified. In these cases we use a more informative prior for θ, $\theta \sim \text{Gamma}(10, 1)$, and still obtain reasonable estimates. Comparing the simulation results of both MCMC methods we find similar point estimates, but slightly higher variability of the estimates of method 2.

To compare both methods with respect to their efficiency Table 1 shows the effective sample size (ESS) for the parameters π, β_0 and θ. The global impression is that both methods show similar performance. However, if the degree of Poisson overdispersion is considered we observe a different behaviour of the methods. Settings a, e, i, and j have the largest values of ϕ within each group of settings with the same E(P), and here method 1 performs better. In all other settings, where ϕ is smaller, method 2 is somewhat better. With respect to computation time method 2 (362 iterations/second) is more than twice as fast as method 1 (150 iterations/second).

4 Application to real data

The model is applied to two examples of count data. First, we consider heart attack counts for males in Styria. The data are monthly counts of hospitalizations after a heart attack for a period of eight years. The motivation for analyzing these data comes from medical experts who expressed belief that not all heart attacks are seen in hospital. There are two reasons for this hypothesis: severe heart attacks are fatal and hence never seen in a hospital, and slight heart attacks may be ignored by the patients. We use a polynomial trend in a strictly parametric regression model for λ_t.

Both methods are applied for estimation yielding very similar results. In fact the difference between the estimates is so small that it is not visible in a graphic like Figure 1, which shows the data, the estimated mean and the estimated total number of cases. The estimated reporting probability is $\hat{\pi} = 0.855$ for method 1 and $\hat{\pi} = 0.851$ for method 2.

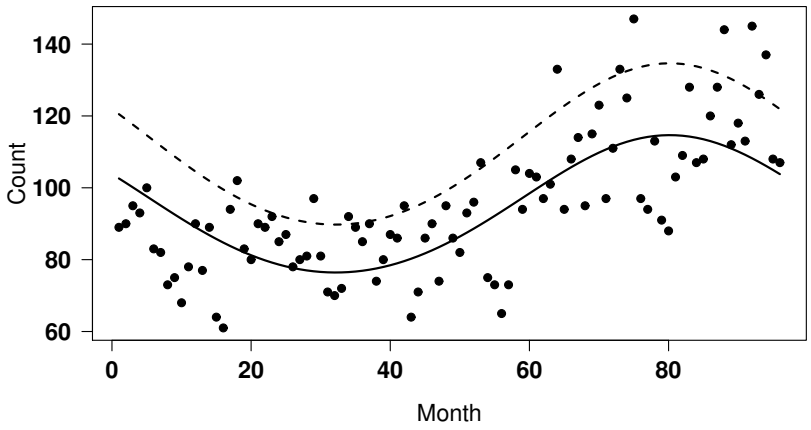


FIGURE 1. Heart attack data, estimated trend for observed (solid line) and total number of cases (dashed line).

The second example are weekly counts of pickpockets in an Austrian city over a period of more than four years. To estimate the trend in an exploratory manner we use the regression model $\log(\lambda_t) = \beta_0 + f(t)$, where $f(t)$ is a P-spline smoother with a large basis ($m = 30$), and apply estimation method 1. Figure 2 shows the data and the estimated trends. Obviously, there is a cyclic component in the trend which is not strictly periodic. The estimated reporting probability is $\hat{\pi} = 0.565$ and hence, roughly only half the cases are reported to the police.

5 Conclusion

In this paper we propose a Bayesian framework for the estimation of underreporting models, and investigate one of them - the beta-Poisson model - in a simulation study. Finally the approach is applied to real data.
Two estimation methods for the beta-Poisson model are suggested, one making use of improved auxiliary mixture sampling and a pure MH algorithm. Comparing both methods in a simulation study we find similar point estimates but differences regarding the effective sample size. Sampling is more efficient using method 1 in cases with high Poisson overdispersion.

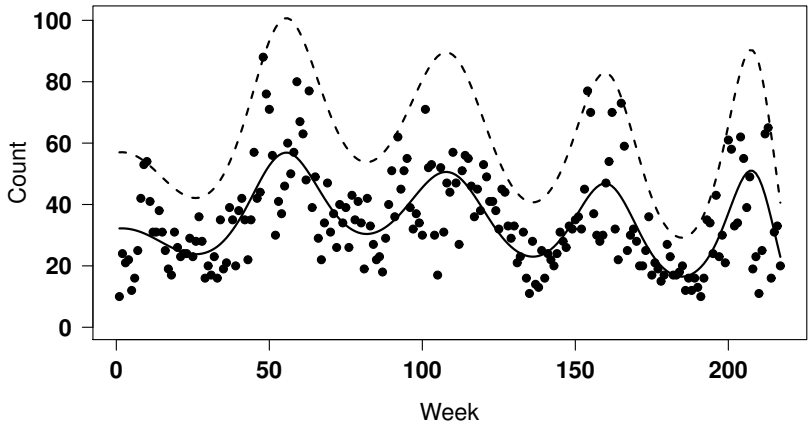


FIGURE 2. Pickpocket data, estimated trend for observed (solid line) and total number of cases (dashed line).

However, in the remaining settings the effective sample size is somewhat higher for the MH method. The improved auxiliary mixture sampler avoids time-consuming tuning of a further parameter, at the expense of computation time, which is more than twice the time of the pure MH method. Applying the methods to real data we find again a good accordance of the two estimation methods, which supports the results from the simulation study.

References

Brezger, A. and Lang, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics and Data Analysis*, 50, 967-991.

Eilers, P. and Marx, B. (1996). Flexible Smoothing with B-splines and Penalties. *Statistical Science*, 11, 89-121.

Frühwirth-Schnatter, S., Frühwirth, R., Held, L. and Rue, H. (2009). Improved Auxiliary Mixture Sampling for Hierarchical Models of Non-Gaussian Data. *Statistics and Computing*, 19, 479-492.

Neubauer, G., Djuraš, G. and Friedl, H. (2010). Models for Underreporting: A Bernoulli Sampling Approach for Reported Counts. *Proceedings of the 9th International Conference: Computer Data Analysis and Modeling*. Minsk.

Expectile contours and data depth

Paul H. C. Eilers[1]

[1] Department of Biostatistics, Erasmus MC, Rotterdam, The Netherlands

Abstract: Data depth is usually based on quantile ideas. Quantiles can be computed by minimizing an asymmetrically weighted sum of absolute values. Expectiles follow from an asymmetrically weighted sum of squares. Convex contours can be constructed by a series of projections of a two-dimensional data cloud and computation of expectiles. A new definition of data depth follows from this construction.

Keywords: Ranks, convex contours, quantiles

1 Introduction

Starting from quantiles of a univariate distribution, one can define the "depth" of a data point. Let $F(x)$ be the cumulative distribution and let $p_i = F(x_i)$ for a (real or potential) observation x_i. Then $d_i = \min(p_i, 1-p_i)$ is called its depth. Intuitively this makes sense: observations with low p and those with high p are found near the extremes of the domain of x. The "deepest" possible x is the median of the distribution, with $d = 0.5$. For an observed univariate data set one can easily compute the depth of each observation using the (discontinuous) empirical distribution.

There is no unique way to define multivariate data depth, where I limit the discussion here to two dimensions. One choice is to compute he convex hull, peel it off like the outer layer of an onion and repeat the process until no data are left. The number of the layer in which a data point occurs (divided by the total number of layers) gives a rough measure of data depth.

Another definition is Tukey depth or half-space depth of a data point (Tukey, 1975). A line through the point splits n observations into two groups. Let the number be $n_1(\phi)$ and $n_2(\phi)$, where ϕ is the angle between the line and a references axis. Then the minimum over all ϕ ($0 \leq \phi < \pi$) of $\min(n_1(\phi), n_2(\phi))/n$ is the Tukey depth. A possible interpretation is that one searches for the quantile with the lowest possible p, for all possible projection angles. Working directly from the definition, computation time for data depth is at least proportional to n^2.

In this paper I propose two changes to the Tukey algorithm. The first step is to choose a fixed grid of angles, project the data points and compute quantiles. Perpendicular lines through the quantiles of the projections define convex regions. This idea is not original: in a recent paper, Hallin et

al. (2009) point to unpublished work by Kong and Mizera (2008). The second step, which I believe to be new, is to replace quantiles by expectiles. It simplifies the computations and it leads to unique and visually pleasing results.

2 Theory and applications

Instead of considering all possible lines through one point and the remaining ones, let us project all data pairs $(x,\ y)$ onto a line at an angle ϕ and compute data depth of the projections. Repeating this process on a grid of values for ϕ, a list of depths is obtained for each observation. The minimum in the list is taken as the final depth. Now computation time is proportional to the product of the size of the grid for ϕ and n.

It is well known that one can define the pth quantile as minimizers of an asymmetrically weighted sum of absolute values $\sum w_i|x_i-q|$, with $w_i = p$ if $x_i > q$ and $w_i = 1-p$ otherwise (Koenker and Bassett, 1978). A less familiar concept is the ath *expectile*, which is the minimizer of $\sum w_i(x_i - u)^2$, with $w_i = a$ if $x_i > u$ and $w_i = 1-a$ otherwise. Here $0 < a < 1$ is the asymmetry parameter which plays the same role as p for quantiles. This is the most simple application of least asymmetrically weighted least squares (LAWS), proposed as asymmetric least squares by Newey and Powell (1987) and Eilers (1987). See also the paper by Schnabel and Eilers (2009) for a recent application of expectiles in smoothing.

For a sample, quantiles are not uniquely defined. It is common to set $\hat{F}(x_{(i)}) = i/n$ if $x(i)$ is observation i after sorting in ascending order, and to interpolate linearly between (sorted) observation, but other choices are possible.

For any theoretical distribution, as well as for any univariate data set, one can compute $a(x)$, which I call the asymmetry function. By analogy we can define $\delta = \min(a, 1-a)$ as data depth. The empirical asymmetry function for a sample is continuous and unique, also between the observations, so no additional interpolation recipe is needed. Interestingly, according to this definition the mean has largest depth, 0.5.

We have a scatterplot of data points $(x_i,\ y_i)$, for $i = 1, \ldots, m$. We draw a line through an arbitrary point, e.g. through $(\bar{x},\ \bar{y})$, at an angle ϕ and project the data points on that line. Let $z_i = x_i \cos\phi + y_i \sin\phi$ be the projection of data point i. For any asymmetry a we can compute the corresponding expectile $u(a)$ of z. Now draw another line, l, perpendicular to the first line, at the position $u(a)$. Increase ϕ to get ϕ' to get a second line, l', and compute the point where l and l' cross. Repeat until a full circle has been completed. The set of crossing points defines a convex region. We call this the region with expectile data depth (EDD) a. Figure 1 shows a set of lines and the region they define. Repeating the above procedure for different values of a, we get a set of nested convex regions, each having EDD a.

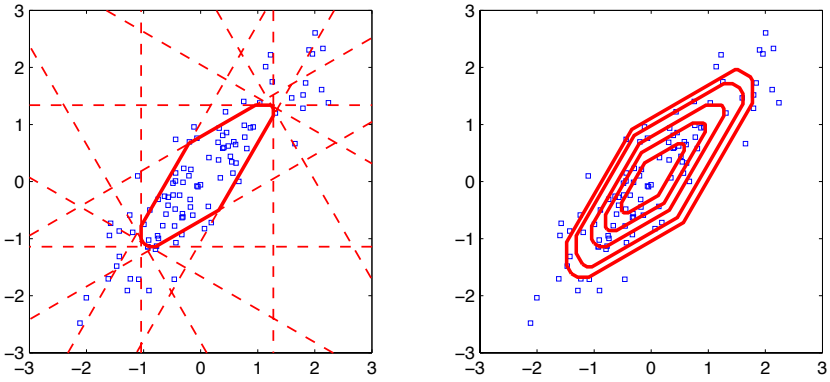


FIGURE 1. Illustration of the construction of expectile contours. Left: individual lines (broken) through $a = 0.05$ expectiles for projection angles that change in steps of 30 degrees, and the convex contour they define (thick line). Right: convex contours for $a = 0.01$, 0.02, 0.05, 0.1 and 0.2.

With quantile contours this simple sequential computation of the corner points of contours will not work, especially for small steps of ϕ and p near 0.5, as experience shows. The root of the problem is the non-uniqueness of sample quantiles. The set of perpendiculars defines a convex inner region, but it is not easy to compute the corner points. The number of corner points can be (much) less than the number of projection angles. A number of intersections of perpendiculars for adjacent projection angles lie outside the convex inner region. This problem does not occur with expectile contours. I have no proof of this, but I did not meet any counterexamples yet. Figure 2 illustrates the problem with quantile contours: we easily see the inner convex region, but to find the corner points one cannot simply let the lines intersect in the order in which they are computed. For expectiles this simple recipe does work.

The computation of the intersection points is very simple. All points on the perpendicular through an expectile u of the projections with angle ϕ have the property that $x \cos\phi + y \sin\phi = u$. To find the intersection with the perpendicular for angle ϕ', through expectile u', we solve the linear system of equations

$$\begin{pmatrix} \cos\phi & \sin\phi \\ \cos\phi' & \sin\phi' \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} u \\ u' \end{pmatrix}. \tag{1}$$

The solution for x is $x = (u \sin\phi' - u' \sin\phi)/\sin(\phi - \phi')$. Likewise $y = (u' \cos\phi - u \cos\phi')/\sin(\phi - \phi')$, where I have used the fact that $\cos\phi \sin\phi' - \cos\phi' \sin\phi) = \sin(\phi - \phi')$. Notice that $\phi - \phi'$ is the constant size of the step between angles on the grid. For visual display one lets ϕ take not too small steps, say 5 or 10 degrees. Or, if one feel that "exactness" is needed, one

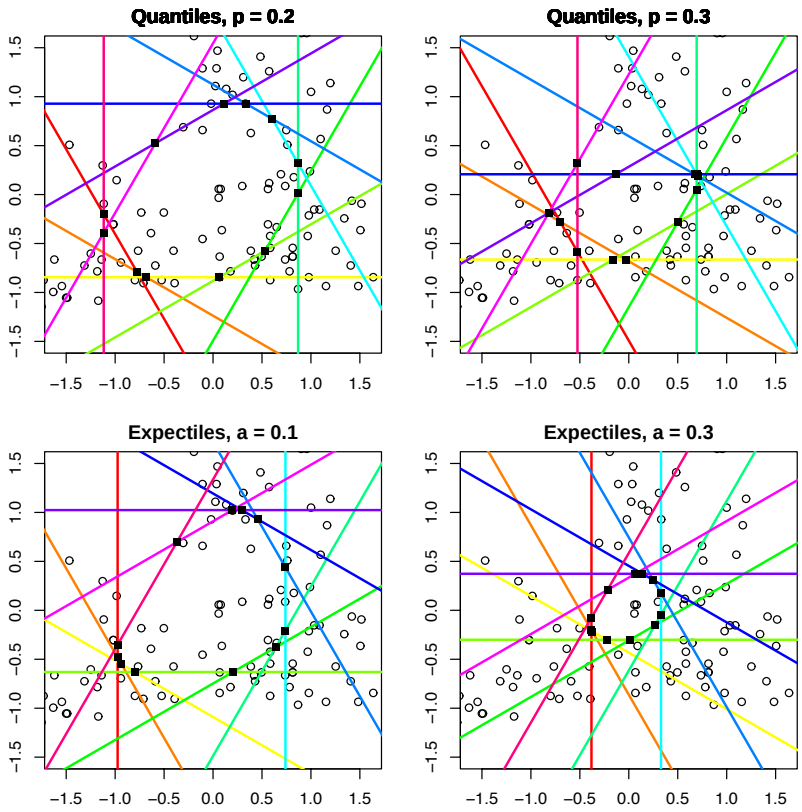


FIGURE 2. Projection perpendiculars for 30 degree steps of the projection angle ϕ. The circles represent the data points. The black squares show the the intersections of each perpendicular with the one computed for the next projection angle. The top row shows two examples for quantiles, the bottom row for expectiles. For quantiles some intersection points lie outside the contour of the inner convex region, but for expectiles they all lie on the contour.

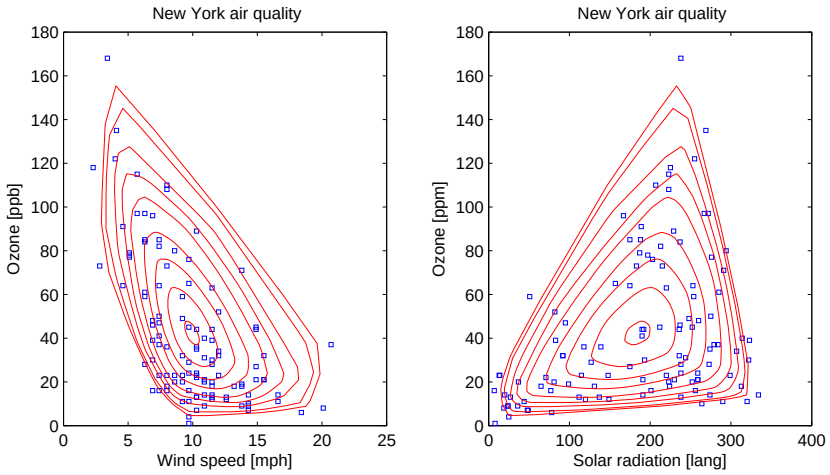


FIGURE 3. Two examples of expectile contours for air pollution data. The asymmetry values are 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, and 0.4.

reduces this to one degree or less, accepting the extra computational load. Computation time is approximately equal to the product of number of angles, the number of asymmetry values and the number of observations. Figure 3 shows an application to the data set `airquality` in the R system.

3 Discussion

Expectile contours are easy to compute and display, and they help to get a better impression of data clouds in two dimensions. In principle extension to three dimensions is straightforward, using a two-dimensional grid of projection angles. For each pair of angles, and for a given asymmetry, a plane is found, perpendicular to the projection lines. Together these planes define a convex polytope. Display of the results is a challenge. A possibility is to let expectile data depth determine color or size of data points.

By construction the contours are convex, so they are less attractive when the data cloud is strongly curved. It is far from clear if generalizations are possible. In any case a more complicated concept than expectiles of projections will be needed.

Expectiles are more sensitive to outliers than quantiles, so the outer contours may be be deformed unacceptably. A possible solution is to apply trimming: remove all data points outside the outermost contour and repeat the computations.

Interestingly, a half-space argument also applies to expectile contours. Any point in the plane, within the convex hull of the data, can be projected on a

line with angle ϕ. From the data follows the asymmetry, a, of the projection. For a certain ϕ, a will achieve its minimum, say $\hat{a}$. The expectile contour corresponding to $\hat{a}$ will run through the point, at an angle perpendicular to the projection line.

References

Eilers, P.H.C. (1987). Asymmetric least squares: new faces of a scatterplot (in Dutch). *Kwantitatieve Methoden* **8**, 45–62.

Hallin, M., Paindaveine, D. and Siman, M. (2010) Multivariate quantiles and multiple-output regression quantiles: from L_1 optimization to halfspace depth. *The Annals of Statistics* **38**, 635–669.

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

Kong L. and Mizera, I. Quantile tomography: using quantiles with multivariate data. arXiv:0805.0056v1.

Newey, W.K. and Powell, J.L. (1987). Asymmetric least squares estimation and testing. *Econometrica* **55**, 819–847.

Schnabel, S.L. and Eilers, P.H.C. (2009) Optimal expectile smoothing. *Computational Statistics and Data Analysis* **53**, 4168–4177.

Tukey, J.W. (1975). Mathematics and the picturing of data. *Proc. Int. Congress Math., Vancouver 2*, 523-531.

Identifying genes under selection using generalized linear mixed models

Kirsten E. Eilertson[1], Jim Booth[2], Carlos D. Bustamante[3]

[1] 102 F Weill Hall, Cornell University, Ithaca NY 14853; kee23@cornell.edu
[2] 1178 Comstock Hall, Cornell University, Ithaca NY 14853
[3] 300 Pasteur Drive Lane Building, Room L-301 Stanford University, Stanford CA 94305

Abstract: We present an approach for identifying genes under selection using polymorphism and divergence data from synonymous and nonsynonymous sites within genes. A generalized linear mixed model is used to model the relationship between the number of polymorphic/divergent mutations and synonymous/nonsynonymous sites. Based on the theory behind the McDonald-Kreitman statistic, we use the estimated fixed and random effects of the model to identify genes under positive and negative selection. The model is fit in both the standard and Bayesian settings using the lme4 package in R, and Markov Chain Monte Carlo Methods in WinBUGS. The proposed methodology is designed to analyze several thousand genes from the same phylogeny. Using simulated data we compare our method to existing methods for detecting genes under selection, the MK statistic, and MKprf.

Keywords: Generalized linear mixed models; natural selection; genetics; Laplace approximation.

1 Introduction

The goal of this research is to develop a methodology which combines results from population genetic models which help us identify signatures of selection in DNA, with statistical techniques that take advantage of the vast amount of information now available. There are many reasons to try to find genes under selection. Positive selection could be indicative of adaption or new form or function. Negative selection could be indicative of susceptibility to disease. Purifying selection, or negative selection so strong that there is a significant lack or absence of mutations, could indicate the gene has significant functional importance. The method we propose is in theory robust to demography (population expansions, bottlenecks, migration, etc.) and also takes advantage of having information provided by multiple genes in the same phylogeny to estimate parameters of interest. In simulations this has been the case. Additionally, in simulations, our method has a significant improvement in power over similar tests while maintaining a low false positive rate.

TABLE 1. y_{ij} = the number of mutation a gene has in category ij; $i = 1$ if the mutations are nonsynonymous, 0 otherwise; $j = 1$ if the mutations are divergent, 0 otherwise

	Polymorphic	Divergent
Synonymous	y_{00}	y_{01}
Non-Synonymous	y_{10}	y_{11}

Motivated by the analyses of McDonald and Kreitman (1991) we present a modeling framework for identifying genes under selection using MK table data, see Table 1. The MK statistic makes use of both polymorphism and divergence data by comparing the trends of polymorphism to divergence for nonsynonymous and synonymous mutations by applying Fisher's exact test to the MK table. A mutation that occurs in every individual in the sample from one species is considered divergent (D), otherwise considered polymorphic (P). A mutation that occurs where it changes the amino acid produced is considered nonsynonymous (N), otherwise considered synonymous (S). If the mutations are not under selection, one would expect $DS/PS \approx DN/PN$. The MK statistic is robust to demography because nonsynonymous and synonymous sites are interspersed among each other and should be similarly affected by demography and genetic drift.

Our method models MK table data using generalized linear mixed models. In this setting, we incorporate genome wide effects into our analysis as fixed effects, and individual gene effects as random effects, thus allowing us to pool information across genes (individual MK tables) and increase power to detect genes under selection. Like the MK test, this test is robust to demography. The test cannot detect recent selective sweeps because both nonsynonymous and synonymous mutations linked to the beneficial mutation will be similarly affected by the selective sweep. The test is best designed, then, to detect recurrent directional selection (including older selective sweeps).

Bustamante et al (2002) developed MKprf as a method that directly estimates the posterior distributions of genomic parameters such as the species divergence time based on the MK tables' synonymous cell entries and the Poisson random field framework, see Sawyer and Hartl (1992). The posterior of the selection coefficients for each gene are then calculated conditional on these genomic parameters and the nonsynonymous cell entries in the MK table. MKprfK is a variation of this method.

2 Method

Let n be number of genes in the sample. Thus we have $4n$ mutation counts y_{ijk}, where $i = 1$ if the count is of nonsynonymous mutations, 0 otherwise,

$j = 1$ if the count is of divergent mutations, 0 otherwise, and $k = 1, ..., n$ is the gene identification number. The mutation counts are assumed to be Poisson distributed, $y_{ijk} \sim P(\mu_{ijk})$. The log of the expected mutation count is modeled using a generalized linear mixed effects model. The fixed effects include an intercept, an effect if the mutation is nonsynonymous, an effect if the mutation is divergent, an interaction between these effects. Additionally the model includes four random effects: a gene effect, and the two-way and three-way interactions between the gene, nonsynonymous, and divergent effects. The random effects are fit allowing a general correlation structure. Additionally, an offset term is used to control for the number of sites sampled in the gene where a mutation of type i could occur, $Tsites_0$ for synonymous mutations, $Tsites_1$ for nonsynonymous mutations.

$$log(\mu_{ijk}) = \log(Tsites_i) + \beta + \beta^N i + \beta^D j + \beta^{ND} ij + \beta_k^G + \beta_k^{NG} i + \beta_k^{DG} j + \beta_k^{NDG} ij, \quad (1)$$

The model is fit in R (GLMM) using the lme4 package, and a Bayesian GLMM (B GLMM) is also fit using WinBUGS, see Lunn et al (2000).
Because selection causes nonsynonymous mutations to become fixed within a population at a higer or lower rate than synonymous mutations, evidence of selection can be found in looking at what we term the 'selection effect', $\beta^{ND} + \beta_k^{NDG}$. In the Bayesian framework we construct credible intervals for the selection effect based on the MCMC samples. In the frequentist framework confidence intervals are constructed for the BLUPs using the Laplace approximation, see De Bruijn (1981), to estimate the marginal posteriors for the random effects. If a gene's selection effect confidence interval does not contain zero, the gene is classified as under selection.
Data was simulated using PRFREQ, see Boyko et al (2008). Selection coefficients for our simulations are drawn from various distributions see Figure 1. From our simulations, we see the GLMM method is a significant improvement over other methods, especially when table counts are low, as with a human-like mutation rate of $\theta = .001$, see Table 2. Specifically, the GLMM methods are more sensitive for small (close to zero) and more accurate for extreme valued selection coefficients see Figures 2 and 3.

TABLE 2. Proportion of genes correctly classified under selection where the selection coefficients are from distributions 1, 2 and 3; mutation rate θ

	$\theta = .01$ (Drosophila)			$\theta = .001$ (Human)		
	Dist 1	Dist 2	Dist 3	Dist 1	Dist 2	Dist 3
GLMM	0.95	0.94	0.95	0.85	0.78	0.87
B GLMM	0.94	0.89	0.94	0.86	0.79	0.86
MKprf	0.90	0.84	0.90	0.55	0.46	0.60
MKprfK	0.89	0.84	0.89	0.61	0.31	0.78
MK	0.79	0.68	0.78	0.05	0.00	0.00

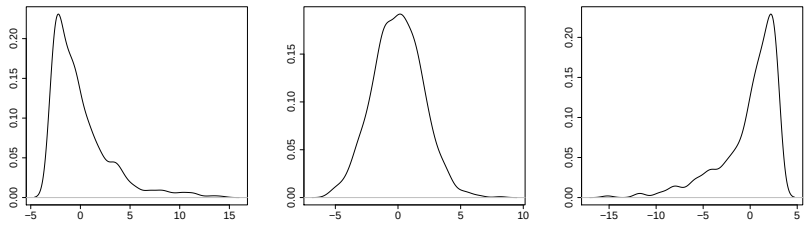


FIGURE 1. Distributions 1, 2 and 3 of simulated selection coefficients.

3 Discussion

This paper focuses on simply identifying genes under selection. However, under additional assumptions such that the Poisson random field (PRF) framework holds, we can set the predicted MK table counts from the GLMM method equal to the expected MK table counts under the PRF framework, and get estimates for the selection coefficient as well as other parameters of interest: the time to the most recent common ancestor τ, mutatation rate θ, and proportion of lethal mutations $1 - f_0$. Simulations show our method to be quite accurate in estimating these parameters, as well as fairly robust to the assumptions of the PRF framework. We hope to publish these findings soon.

References

Boyko, A.R., et al (2008). Assessing the Evolutionary Impact of Amino Acid Mutations in the Human Genome. *PLoS Genetics* **4**

Bustamante, C.D., Nielsen, R., Sawyer, S.A., Olsen, K.M., Purugganan, M.D. and Hartl, D.L. (2002). The cost of inbreeding in Arabidopsis. *Nature* **416** 531 - 534

De Bruijn, NG (1981). *Asymptotic methods in analysis*. Dover Pubns.

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

McDonald, J.H. and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, **351**, 652 - 654.

Sawyer, S.A. and Harl, D.L. (1992). Population genetics of polymorphism and divergence. *Genetics* **132** 1161 - 1176

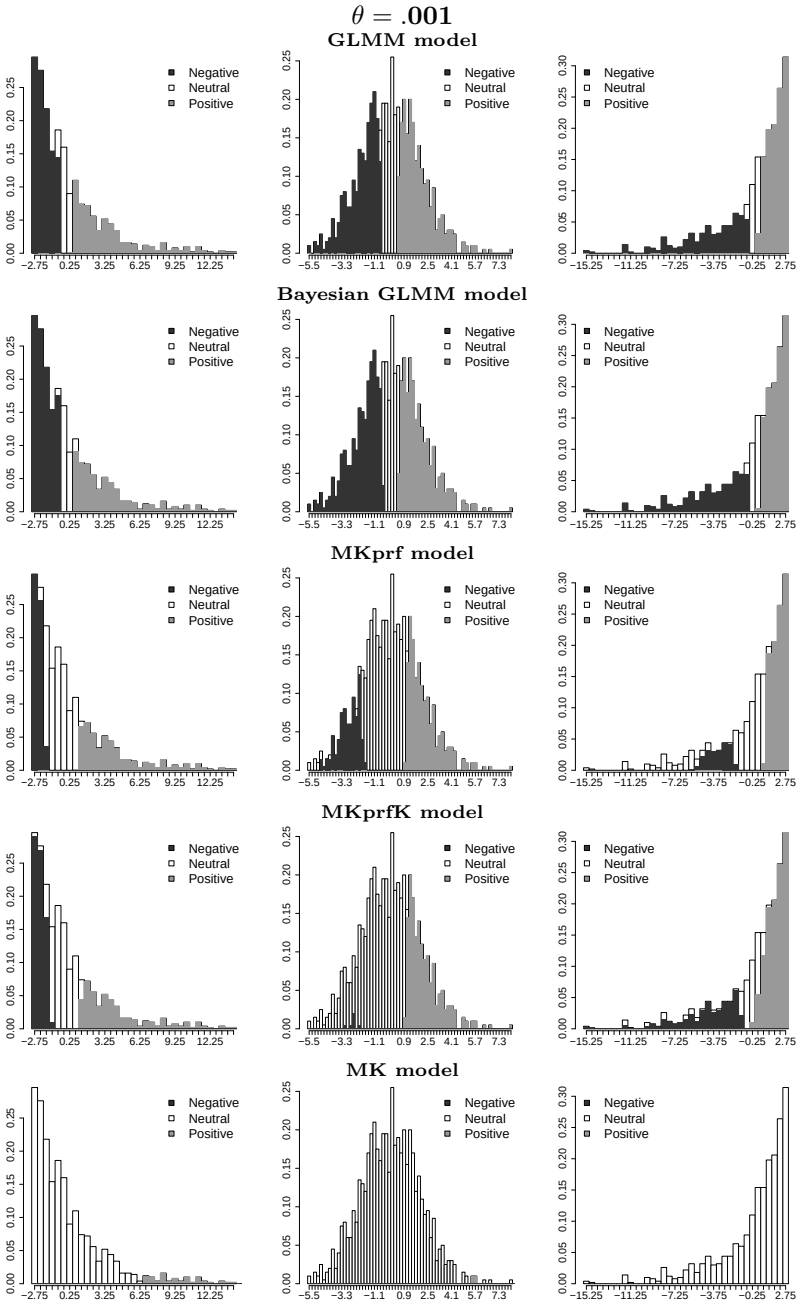


FIGURE 2. Shaded regions of histogram represent the proportion of genes under selection classified as under selection (x-axis is selection coefficient) .

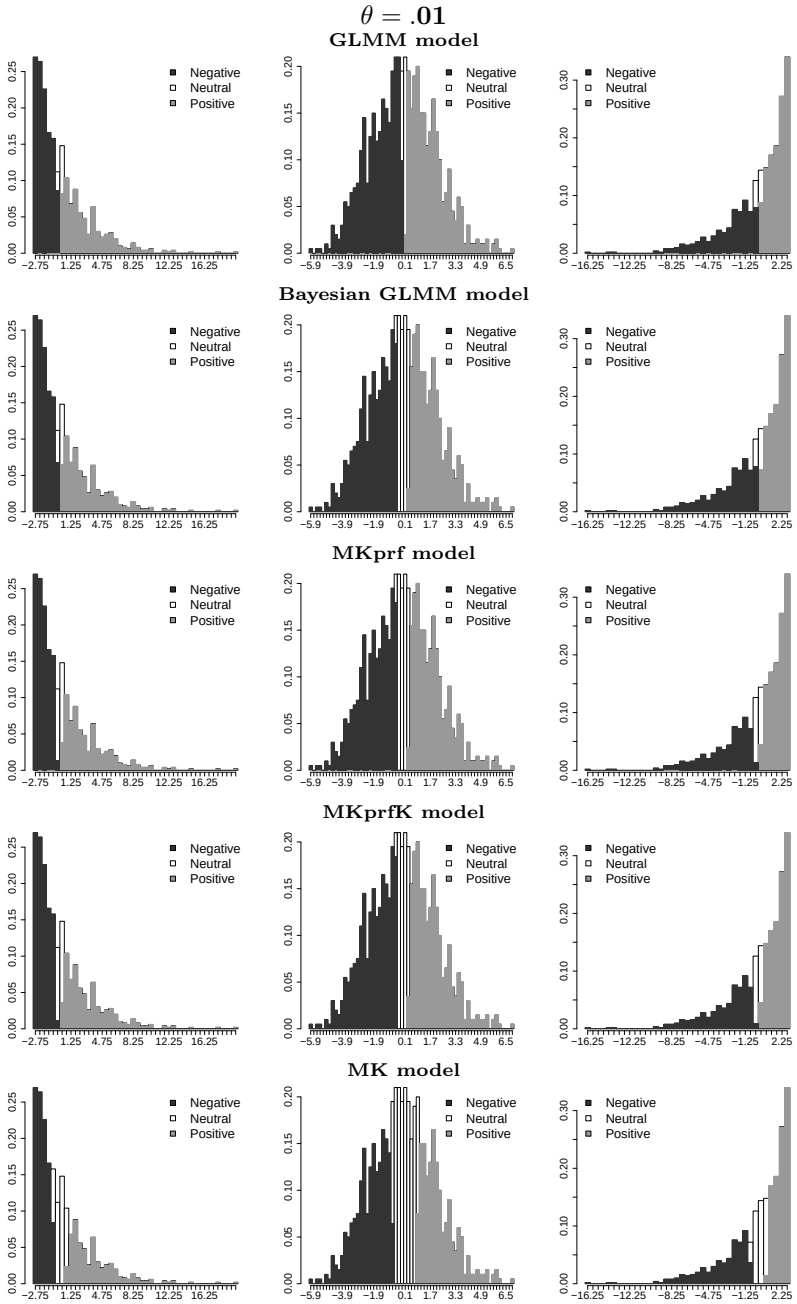


FIGURE 3. Shaded regions of histogram represent the proportion of genes under selection classified as under selection (x-axis is selection coefficient) .

Localized regression on principal manifolds

Jochen Einbeck[1], Ludger Evers[2]

[1] Department of Mathematical Sciences, Durham University, Durham DH1 3LE, England, jochen.einbeck@durham.ac.uk
[2] Department of Statistics, University of Glasgow, Glasgow G12 8QQ, Scotland

Abstract: We consider nonparametric dimension reduction techniques for multivariate regression problems in which the variables constituting the predictor space are strongly nonlinearly related. Specifically, the predictor space is approximated via "local" principal manifolds, based on which a kernel regression is carried out.

Keywords: Smoothing; principal curves and surfaces; localized PCA.

1 Introduction

This article deals with the problem of multivariate regression for situations where the (possibly high-dimensional) predictor space features complex dependency patterns. As an example, consider oceanographic data extracted from the World Ocean Database, which include measurements on the water temperature (serving as the response variable, Y), and the three covariates X_1=salinity, X_2=water depth, and X_3=oxygen content (Fig. 1 left). Obviously, the three covariates are highly and nonlinearly related and contain partially redundant information. Potential modelling strategies include a full interaction model $Y = m(X_1, X_2, X_3) + \epsilon$, which becomes more difficult the more covariates are involved, or an additive model $Y = m(X_1) + m(X_2) + m(X_3) + \epsilon$, which ignores the interaction between the variables.

Neither of these methods exploits the fact that the covariates occupy a space of lower intrinsic dimensionality than 3. Formulating the problem more generally: We are given a regression problem with response Y and predictor space $X = (X_1, \ldots, X_p)^T$. We aim for a two-step strategy which would (1) approximate X nonparametrically by a curve, surface, or, more generally, a low-dimensional manifold of dimension $d < p$, and (2) use the compressed data as a $d-$dimensional predictor henceforth. In this sense, this article provides an extension of principal component regression, being nonparametric both in the compression and the regression step. We assume that the intrinsic dimensionality of the manifold, d, is given, e.g. from visual inspection of the data cloud. Dimensionality estimation is beyond the scope of this paper; an overview on such methods is given in Camastra (2003).

2 Methodology

2.1 The case $d = 1$

We are given independent replicates $x_1, \ldots, x_n \in \mathbb{R}^p$ drawn from the random vector X, i.e. $x_i = (x_{i1}, \ldots, x_{ip})^T$. For the compression step (1), we use the local principal curve algorithm (LPC; Einbeck et al., 2005), which can be summarized as follows. Let w_i^x denote an appropriate (bell-shaped) weight function centered at $x \in \mathbb{R}^p$. Beginning at some starting point $x = x_0 \in \mathbb{R}^p$, we calculate $\mu^x = \sum_{i=1}^n w_i^x x_i$, and then iterate

(i) Compute the first local eigenvector γ^x of $\Sigma^x = (\sigma_{jk}^x)_{(1 \le j,k \le p)}$, where $\sigma_{jk}^x = \sum_{i=1}^n w_i^x (x_{ij} - \mu_j^x)(x_{ik} - \mu_k^x)$ and μ_j^x denotes the $j-$th component of μ^x. Using a step size z, step from μ^x to $x := \mu^x + z\gamma^x$;

(ii) Calculate the local center of mass μ^x;

until the distance between neighboring values of μ^x becomes negligible. The resulting series of μ^x, which defines the local principal curve, is subsequently connected through a cubic spline and parametrized by its arc length. Each data point is then projected to its nearest point on the curve, and the compressed data correspond to their projection index (PI). This is illustrated in Fig. 1 (left). Details on the parametrization and projection are found in Einbeck et al. (2010). In the regression step (2), we regress the response versus the PIs, using any univariate nonparametric smoother (e.g., local linear). This is illustrated in Fig. 1 (right).

2.2 The case $d \ge 2$

The use of localized principal components in (i) is by no means the only possible option. If we replaced γ^x by the direction of, say, the vector connecting the previous and the current local center of mass, then step (ii) would adjust the principal curve again towards the "middle" of the (local) data distribution. This slightly modified algorithm has, just like the original LPC algorithm, line segments as geometric building blocks in step (i). We exploit this idea for the extension of LPCs to local principal manifolds (LPMs). As the basic building block we will now use a triangle ($d = 2$), tetrahedron ($d = 3$), or simplex ($d \ge 4$). Although the algorithm that we are going to propose can in principle be applied using any $2 \le d < p$, we will describe it for ease of presentation for the special case $d = 2$, in which case the resulting object is a local principal surface (LPS).

Given a triangle Δ on the boundary, we extend the surface by attaching new triangles to its "free" edges. The triangles are obtained by reflecting Δ at the free edges. Suppose that the current triangle Δ has the vertices δ_1, δ_2, and δ_3, and that the edge (δ_2, δ_3) is a free edge beyond which we want to extend the surface:

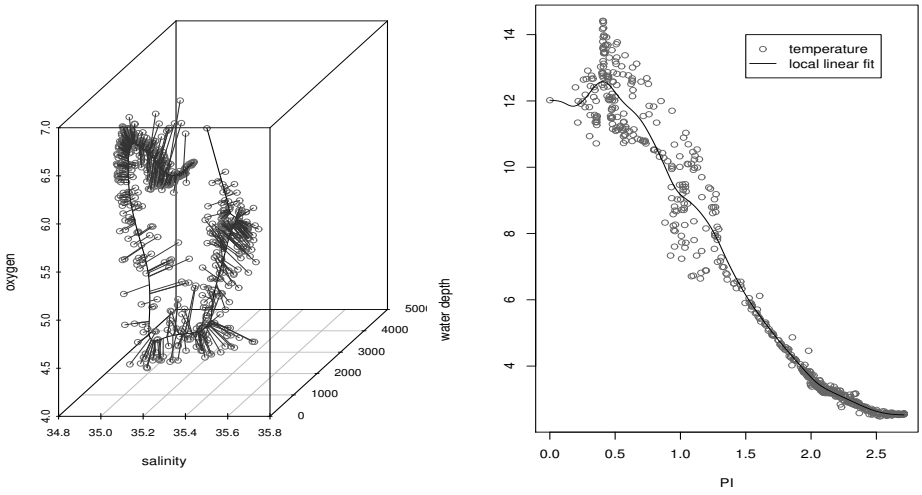


FIGURE 1. Left: 3d- scatterplot (grey circles) of salinity (measured on the 'Practical Salinity Scale'), water depth (metres), and oxygen content (millilitre/litre of water). Solid curve: cubic spline representation of local principal curve, with orthogonal projections; right: water temperatures plotted vs. projection indices.

(i) A preliminary vertex $\tilde{\delta}_4$ is obtained by attaching an equilateral triangle to the edge (δ_2, δ_3) such that δ_1, δ_2, δ_3, and $\tilde{\delta}_4$ all lie on the same plane. The bottom right point in Fig. 2 (left) illustrates this preliminary vertex.

(ii) Compute δ_4 from $\tilde{\delta}_4$ as a constrained local center of mass, which enforces that the triangle with vertices δ_2, δ_3, and δ_4 is equilateral. Fig. 2 (left) shows the weights of the observations (darker grey corresponds to higher weights), with the circle representing the constraint. The new vertex δ_4 is shown in the top right. The newly-created triangle is dismissed if an already existing vertex lies in its circumsphere or if the new vertex δ_4 lies in the circumsphere of an existing triangle (in the former case δ_4 is replaced by the already existing offending vertex), or if the new vertex falls into a region of small density.

The initial triangle is placed in the plane spanned by the first two local principal components obtained at a (manually or randomly chosen) starting value x_0. Steps (i) and (ii) correspond to their counterparts in the LPC algorithm. The checks for dismissal of vertices in (ii) ensure that branching triangles "meet" again and do not form many parallel surfaces.
We now apply the LPS algorithm to the oceanographic data. The fitted surface, which features 177 triangles with an average count of 3.63 data

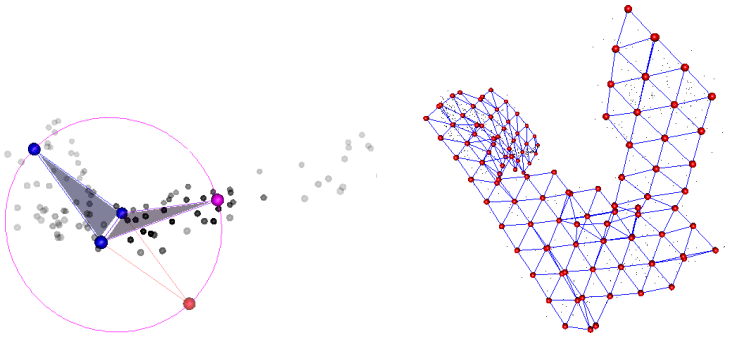


FIGURE 2. Left: Illustration of the LPS algorithm; right: Fitted LPS for oceanographic data.

points per triangle, is shown in Fig. 2 (right): it nicely captures the shape of the data cloud.

It is hard to find a full bivariate parametrization of the LPS. Therefore, we use a simple kernel regression. For each pair of triangles we define the (discrete) "distance" d as the smallest number of triangle borders that need to be crossed to proceed from one triangle on the surface to the other one. This distance can be obtained by applying Dijkstra's algorithm to the neighborhood graph, and is thus cheap to compute. In order to assign local weights, we define the discrete distance-based kernel $\kappa(d) = e^{-d/\lambda}$, where λ is a smoothing parameter. Special cases are $\lambda = 0$, corresponding to no smoothing at all, and $\lambda \longrightarrow \infty$, where the estimated response function is constant. The smoothed response value $\hat{y}_\Delta$ on triangle Δ is then given by

$$\hat{y}_\Delta = \frac{\sum_{\Delta'} \kappa(d_{\Delta,\Delta'})\bar{y}_{\Delta'}}{\sum_{\Delta'} \kappa(d_{\Delta,\Delta'})},$$

where $\bar{y}_{\Delta'}$ is the mean of all observations for which Δ' is the closest triangle, and $d_{\Delta,\Delta'}$ is the discrete distance between the triangles Δ and Δ'. Though formulated here in the special case $d = 2$, both the estimation of the manifold, as well as the kernel regression on it, extend straightforwardly to higher intrinsic dimensions $d > 2$ by using the appropriate geometric building block.

In order to study the performance of this technique, we split the $n = 643$ observations into a training set of size 500 and a test set of size 143. We include in our study the additive model (AM) as well as localized regression on a local principal curve (LPC) or surface (LPS). The training data are used to learn these nonparametric models. The smoothing parameters for the smooth terms in the additive model and the local smoother on the principal curve are calibrated so that a total of ≈ 16 degrees of freedom is used in each model. For the regression on the surface, we compare three different choices of the smoothing parameter λ. The results of this study

		AM	LPC	LPS		
				$\lambda = 0.2$	$\lambda = 1$	$\lambda = 2$
Training	mean	0.08946	0.32606	0.04335	0.07380	0.14444
error	median	0.01538	0.00650	0.00143	0.00655	0.01455
Test	mean	0.15494	0.30962	0.11090	0.11569	0.17471
error	median	0.02855	0.00877	0.00395	0.01009	0.02059

TABLE 1. Mean and median prediction errors for the training and test data; using AM-, LPC- and LPS-based regression, respectively.

are displayed in Table 1. As expected, the LPC-based regression is inferior to the additive model in terms of the mean prediction error (i.e., the mean of squared distances between predicted and true temperature). The poor performance of the LPC-based technique is due to the branched shape of the response data seen in Fig. 1 (right). The LPS-based approach clearly outperforms the additive model for $\lambda \leq 1$, though for $\lambda = 0.2$ considerable overfitting (undersmoothing) appears to be present, which is reflected in test errors that are about three times larger than the training errors. The choice $\lambda = 2$ leads to larger prediction errors; here we have oversmoothed. Considering the *median* instead of the mean prediction error, the performance of all investigated methods improves drastically (relative to the additive model), which can be explained with an increased robustness of the median to very poor predictions, which can occasionally happen for the LPC/LPM- based approaches especially in the boundary regions.

3 Conclusion

We have presented an entirely nonparametric approach to modelling data which feature a low-dimensional non-linear latent structure. Just like the local principal curves (LPC) algorithm, this local principal manifolds algorithm (LPM) is based on the simple geometric idea of locally approximating the data by connected simplices.

Of course, not every data set will have such a low-dimensional structure. The majority of data sets probably do not, but there are still surprisingly many datasets which do have such a structure. Once the algorithm has established the low-dimensional latent structure, one can use it to define new, data-dependent topologies, which often give a better representation of the dynamics underlying the data than the standard Euclidean distance in the original data space. This implied dimension reduction can, for example, be exploited when studying regression problems, as illustrated in the example shown in the preceding section. Other applications include classification or density estimation on the manifold.

Acknowledgments: We wish to thank Benedict Powell, Durham University, for retrieving, preparing, and sharing his insight into the oceanographic data set.

References

Camastra, F. (2003). Data dimensionality estimation methods: a survey. *Pattern recognition* **36**, 2945–2954.

Einbeck, J., Tutz, G., and Evers, L. (2005). Local principal curves. *Statistics and Computing* **15**, 301–313.

Einbeck, J., Evers, L., and Hinchliff, K. (2010). Data compression and regression based on local principal curves. In Fink et al. (Eds): Advances in Data Analysis, Data Handling, and Business Intelligence, pp. 701–712, Heidelberg: Springer.

Modelling Alkalinity in Ecosystems

Jude Eze[1], E. Marian Scott[1], Adrian Bowman[1], Mark Hallard[2], Claire Ferguson[1], Duncan Lee[1]

[1] Department of Statistics, 15 University Gardens, University of Glasgow, G12 8QW

[2] Scottish Environment Protection Agency. Carseview House, Castle Business Park, Stirling FK9 4SW

Abstract: The assessment of surface water quality can provide valuable information on the anthropogenic, biological and chemical activities and the geomorphology of the surrounding area. It is therefore of interest to analyze the temporal and spatial trends of determinands in surface waters that have direct consequence on the quality and life of ecosystems and biodiversity. In particular, the level of alkalinity is an important factor in the survival of different aquatic and plant species and hence the conservation of ecosystems. The trend in the buffering capacity of water bodies in the Loch Lomond and the Trossachs National Park are analyzed and variation of this capacity between different locations, where monitoring data exist, are examined. Results indicate the existence of significant temporal trends, seasonality and spatial pattern.

Keywords: Time Series Analysis;Spatial variation; Ecosystems.

1 Introduction

The physical and chemical features of surface water bodies to some extent reflect the geology, physical and biochemical activities in the surrounding areas. Therefore, water quality is an important indicator of catchment activities and monitoring its trend may give information on possible contaminants, health and wellbeing of the ecosystem (WFD 2000).
Our interest is to analyze the trend, seasonality and spatial variation of alkalinity, measured by the level of $CaCO_3$ in the water bodies within the Loch Lomond and the Trossachs National Park (LLTNP). This determinand is chosen based on its strategic importance for the conservation of plant and aquatic life and provision of ecological services in the Park.
Low levels of alkalinity may lead to fish death and very high levels could affect plant life. Therefore, alkalinity is important in the determination of the composition of some biological communities and useful in setting ecological quality metrics. The Freshwater Fish Directive (FWFD) classified suitable waters for different fish species based on alkalinity and altitude (FWFD 1978, UKTAG, 2006).

Spatial variation in alkalinity depends greatly on the geology of an area, waste water effluents, acid rain (in rainfall), carbonate rich soil and land use.

1.1 Data

There are 22 large lochs and about 50 rivers and large burns in the Loch Lomond and the Trossachs National Park (LLTNP, 2005). The Scottish Environment Protection Agency (SEPA) has about 289 water quality monitoring sites where data on more than 108 determinands are routinely collected. Most of the sites have data collected approximately monthly from 1987 to 2009. However, only 170 of these sites monitor alkalinity. A 12 year period, from 1998 to 2009 was selected. Selection of sites for analysis was based on two criteria, data availability and number of observations. Forty-three sites with at least 40 observations measured between 1998 and 2009 were selected.

2 Methods

Data for each of the 43 sites were analyzed separately to enable comparison between sites. To reduce the effect of outliers and stabilize the variance across time, we applied a natural log transform of the data. An additive model was then fitted to the log alkalinity, modeled as a sum of smooth basis spline functions of day of the year, year (containing the decimal day of the year), and error terms, where $error \sim N(0, \sigma^2)$, which at this stage are assumed to be independent. For seasonality, cyclic smoothers were used for the day of the year. Analysis was conducted using the *mgcv* package in *R* (Wood, 2006).

$$ln(alk) = s(day.site) + s(year) + error \tag{1}$$

We extended this model to account for variation of alkalinity over space. At this stage however, we assume that there are no interactions between time and space and that errors are independent.

$$ln(alk) = s(day.site) + s(year) + s(Easting, Northing) + error \tag{2}$$

3 Results

Analysis of all the 43 sites indicates that alkalinity is generally higher among river sites compared to the lochs. Figure 1 displays plots of estimates and original data (in log units), seasonality and trend pattern obtained from model (1) for a selection sites. Results from model (1) confirm

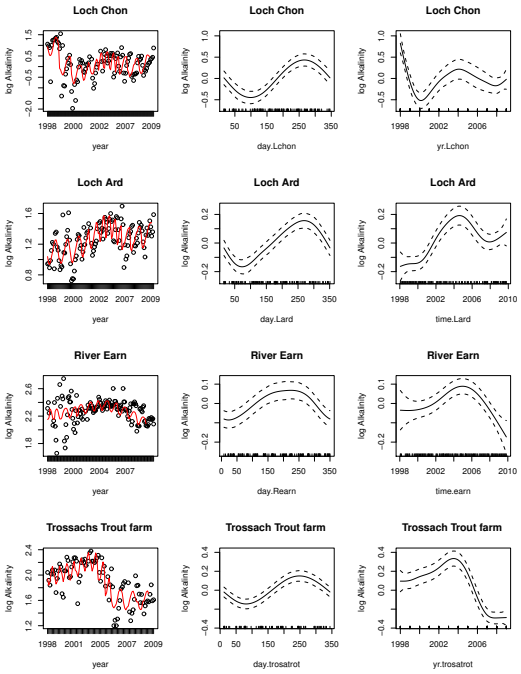


FIGURE 1. Plot of results from model (1) showing from left to right, estimates (in red) and original data (in log units) for each site , the day (seasonal) and year (trend) effects represented by solid lines. The dashed lines are the standard error bounds.

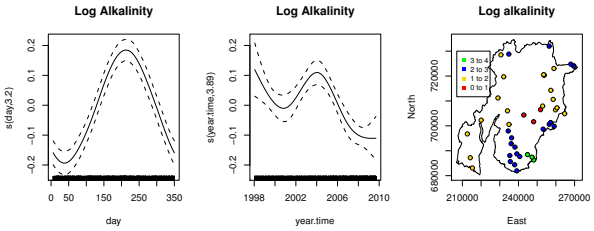


FIGURE 2. Results from the spatiotemporal model (2). Showing from left to right, the plot of seasonality, temporal trend and the spatial pattern at a fixed time point.

statistically significant temporal trends and seasonality in most sites. Lower levels of alkalinity were observed in winter and spring with peaks in the summer. Alkalinity was unusually high in 2004 in most sites and the general pattern can be described as nonlinear with some sites showing signs of continuous decline since 2004 which may have implications for the fish population in those sites.
Analysis using model (2) also confirms statistically significant seasonality, temporal trend and spatial effects. The plots in Figure 2 show from left to right, the estimates of seasonality, trend and spatial pattern. Sites like Lochs Chon, Ard and Achray (represented by red dots) in the central part of the park have the lowest concentration of alkalinity while the highest levels are found among sites in the south (see Endrick river and Cater burn represented with green dots). The overall trend as shown in the middle plot of Figure 2 suggests a decline in alkalinity since 2004.

4 Work in progress

The relationships between alkalinity and other covariates are being investigated with a view to establishing the drivers that induce changes in alkalinity levels. We have so far assumed independence over time and space. The next stage will examine the existence of interaction between time and space and if errors are correlated.

Acknowledgments: Special thanks to NERC for funding this research and to SEPA for providing the required data

References

Lock Lomond and the Trossachs National Park . State of the Park Report 2005

FWFD (1978). Council Directive 78/659/EEC on the quality of fresh waters needing protection or improvement in order to support fish life (OJ L 222, 14.8.78)

UKTAG Report (2006). UK Environmental Standards and Conditions Final Report `http://www.wfduk.org/UK-Environmental-Standards/-LibraryPublicDocs/`. Accessed 3rd Feb 2010

WFD (2000). Directive of the European Parliament and of the Council 2000/60/EC. Establishing a Framework for Community Action in the Field of water Policy, European Union, Luxembourg.

Wood S. N. (2006). *Generalized Additive Models . An Introduction with R.* Chapman & Hall/CRC

Small area estimation for a latent variable: the case of disability in the Italian National Health Interview Survey

E. Fabrizi[1], G. E. Montanari[2], M. G. Ranalli[2]

[1] Dip. di Scienze Economiche e Sociali, Università Cattolica, Piacenza, Italy
[2] Dip. di Economia, Finanza e Statistica, Università degli Studi di Perugia, Italy

Abstract: Quantifying the amount of population in a condition of severe disability that requires intensive care is very important in Italy for its consequences on Health system organization, policy making and funding. To this purpose, only data from the National survey on Health Conditions and Appeal to Medicare can be used, in which, however, no direct measurement of such condition is taken. Fourteen items are available from the questionnaire, which surveys a set of functions concerning the ability of a person to accomplish everyday tasks such as getting washed and dressed, eating and walking. Latent Class Models can then be employed to classify the population according to different levels of a latent variable connected with disability. The survey, however, is designed to provide reliable estimates at the level of Administrative Regions – NUTS2 level. Administrative Regions in Italy are divided into Health Districts and the local Authorities are interested in quantifying the amount of population that belong to each latent class for each District and, possibly, age class. Therefore, small area estimation techniques should be used. The challenge of the present application is that the variable of interest is not observed. We propose to tackle the problem of classifying the population and getting small area estimates as a whole within a Hierarchical Bayesian framework in which the probability of belonging to each latent class changes with covariates. Age by sex by marital status counts are available for each municipality from administrative registers and can be used to this end. The functional form of the influence of age in learnt from the data using penalized splines. A random effect capturing the variability of the small areas is also introduced.

Keywords: Latent Classes; Hierarchical Bayes; Penalized splines; Unit Level Model

1 The problem and the data

Italy has the largest proportion of population aged 65 or more among European countries. Ageing of the population is a central issue for policy makers. In particular, quantifying which proportion of the population has a condition of severe disability that requires intensive care is very important for its consequences on Health system organization, policy making and

funding. To this purpose, data from the national survey on Health Conditions and Appeal to Medicare 2004-2005 conducted by the Italian National Institute of Statistics can be used. The questionnaire is constructed accounting for the International Classification of Impairments, Disabilities and Handicaps developed by the World Health Organization in 1980. Disability is evaluated by means of 14 items in the questionnaire that include the Activities of Daily Living. Four types of disability are defined according to the kind of deprived functional autonomy: confinement, difficulties in movement, difficulties in everyday activities and tasks, sensory deprivation. Table 1 reports the 14 items and their categorization. A condition of permanent constriction in bed, on a chair or at one's home due to physical or psychical reasons is intended for confinement. People with difficulties in movements show problems in walking, i.e. they can only walk few steps before taking a rest; they cannot climb the stairs without stopping; they cannot bend to pick up something from the ground. Difficulties in the activities of daily living are concerned essentially with a lack of independence in accomplishing basic everyday tasks as going to bed, sitting, getting dressed or washed, taking a bath or a shower. Finally, sensory deprivation includes limitations in hearing - e.g. not being able to listen to a TV show even at a high volume; limitations in seeing - e.g. not being able to recognize a friend at a meter distance; limitations in talking. The items are all ordinal with categories increasing with the difficulty of fulfilling the task.

A person is considered disabled in the National Survey if he/she expresses the largest degree of difficulty in at least one of these items. However, the disability that is of interest to the Policy maker is a more severe one that has particular consequences on the care needed. A severely disabled person is affected by a permanent disability or a chronic disease that deprives his/her autonomy so much that he/she needs continuous personal assistance to complete basic everyday tasks. To classify the population according to different levels of disability, Latent Class Models (Lazarsfeld and Henry, 1968) can be employed. Montanari et al. (2009) use this approach to classify people aged 65 or more and are able to identify one class as composed by those in a condition of severe disability. In this work we follow this approach, but we focus on the whole population of the Italian central administrative region of Umbria; here the survey has involved about 1,200 households and the aforementioned items are surveyed on 2,952 people. The survey uses a two stage stratified clustered sampling design and provides direct estimates reliable up to the Administrative Region level (NUTS2). However, Umbria is divided into Health Districts and the local Administrative Department responsible for Health organisation and planning is interested in quantifying the amount of severely disabled population for each District and age class (6-24; 25-64; 65-75; 75 and more). Once a unit in the sample is labelled with his/her most probable latent class, reliable direct estimates of the amount of population within each class cannot be obtained at a subregional level. Small area estimation techniques should

TABLE 1. Items' description and categorization.

Type of disability	Item description	Categories
Confinement	type of confinement	0 = No
		1 = confined to one's home
		2 = confined to a chair
		3 = confined to one's bed
Difficulties in movements	longest walkable distance	0 = More than 200 m.
		1 = Less than 200 m.
		2 = Only few steps
	going up and down the stairs	0 = Yes
		1 = With some effort
		2 = With a lot of effort
		3 = No
	stooping down	Same
Difficulties in everyday activities and tasks	getting in and out of bed	0 = No effort
		1 = With some effort
		2 = With the help of others
	sitting and standing from a chair	Same
	getting dressed and undressed	Same
	taking a bath or a shower	Same
	washing one's face and hands	Same
	eating cutting one's food	Same
	chewing	0 = Yes
		1 = With some effort
		2 = With a lot of effort
		3 = No
Sensory deprivation	hearing a TV show	0 = Yes
		1 = Only at a high volume
		2 = No
	seeing and recognizing a friend	0 = Yes
		1 = Only at short distance
		2 = No
	speaking	0 = Yes
		1 = With some effort
		2 = With a lot of effort
		3 = No

then be used to obtain estimates for the 12 Health Districts by the 4 age classes. The challenge of the present application is that the variable of interest is not observed. We propose to tackle the problem of classifying the population and getting small area estimates as a whole within a Hierarchical Bayesian framework in which the probability of belonging to each latent class changes with covariates. Age by sex by marital status counts are available for each municipality from administrative registers and can be used as covariates. The functional form of the influence of age is learnt from the data using penalized splines. A random effect capturing the variability of the small areas is also introduced.

2 The model

Let Y_{ijt} denote the response of unit i within small area j on item t. The number of small areas is indicated with $J = 12 \times 4$, the number of units for small area j is n_j so that the overall sample is given by $n = \sum_{j=1}^{J} n_j$. The total number of items is $T = 14$. A particular level of item t is denoted by h_t and its number of categories by H_t. The latent class variable is denoted by Q_{ij}, a particular latent class by c and the number of latent classes by C. The full vector of responses of unit i in small area j is denoted by $\mathbf{Y}_{ij}$, whilst $\mathbf{h}$ refers to a possible answer pattern. If N_j is the population size of small area j, we are interested in estimating the small area totals $Q_j(c) = \sum_{i=1}^{N_j} I(Q_{ij} = c)$, for $j = 1, \ldots, J$ and $c = 1, \ldots, C$, where $I(\cdot)$ denotes the indicator function. The latent class small area model can be expressed as

$$\begin{cases} P(\mathbf{Y}_{ij} = \mathbf{h}) & = & \sum_{c=1}^{C} P(Q_{ij} = c) P(\mathbf{Y}_{ij} = \mathbf{h} | Q_{ij} = c), \\ \log \dfrac{P(Q_{ij} = c)}{P(Q_{ij} = C)} & = & U_{jc} + \alpha_c \texttt{sex}_{ij} + f_c(\texttt{age}_{ij}) + \gamma_c \texttt{marital-status}_{ij}, \\ U_{jc} & \sim & N(0, \sigma^2_{Uc}), \text{ for } c = 1, \ldots, C-1. \end{cases}$$

The first equation is the latent class model in which the probability of observing a response pattern $\mathbf{h}$ is a weighted average of class-specific probabilities. In fact, the term $P(\mathbf{Y}_{ij} = \mathbf{h}|Q_{ij} = c)$ is the conditional response probability of observing pattern $\mathbf{h}$ given that unit i in small area j belongs to class c, and the weight is the probability that such unit belongs to the latent class c. By assuming independence of responses within latent classes, the conditional probability takes the form $P(\mathbf{Y}_{ij} = \mathbf{h}|Q_{ij} = c) = \prod_{t=1}^{T} P(Y_{ijt} = h_t|Q_{ij} = c)$. The probabilities $P(Q_{ij} = c)$ are modeled via the multinomial logistic mixed model of the second equation in which the last latent class is the reference one. Age and marital status enter this model parametrically, while $f_c(\cdot)$ is a smooth function of age, for $c = 1, \ldots, C-1$, to be estimated via p-splines. Opsomer et al. (2008) use p-splines for small area estimation for their connection with mixed models and Crainiceanu et al. (2005) offer details for p-splines smoothing within a

Bayesian framework. Finally, random intercepts U_{jc} are included only for the Health Districts to capture the between-areas variation not explained by the covariates. Once prior distributions are assumed on model parameters, a Markov Chain Monte Carlo algorithm can be used to obtain M samples from the joint posterior distribution.
Model selection brings to a good classification in four classes. Looking at posterior probabilities $P(Y_{ijt} = h_t|Q_{ij} = c)$, classes can be interpreted as being composed by people (i) without disability, (ii) with difficulties in movements, (iii) with difficulties in movements and daily tasks, (iv) with severe disability. As of the multinomial model, females are more likely to belong to the last two classes than males, while marital status is not a significant covariate. In addition, the odds of belonging to the first class as opposed to the last one remain constant until the age of about 50 and then decrease steadily with age. Auxiliary information comes in the form of `age` $\times$ `sex` $\times$ `marital-status` population counts for each municiaplity. After dropping marital status from the model, useful auxiliary information is given by `age` $\times$ `sex` classes for each Health District. Marginal MCMC samples can then be used to obtain estimated probabilities $\hat{P}^{(m)}(Q_{\ell j} = c)$, for the ℓ-th class for $\ell = 1, \ldots, L$ and L is given by the number of `age` $\times$ `sex` categories. Estimates of the small area total $Q_j(c)$ can then be computed as

$$\hat{Q}_j(c) = \frac{1}{M} \sum_{m=1}^{M} \sum_{\ell=1}^{L} N_{\ell j} \hat{P}^{(m)}(Q_{\ell j} = c|U_{jc}, \texttt{sex}_{\ell j}, \texttt{age}_{\ell j}),$$

where $N_{\ell j}$ is the amount of population in small area j belonging to class ℓ. Marginal MCMC samples can be used also for variance estimation of such estimated counts (see Malec et al., 1997).

Acknowledgments: Work is partially supported by the project PRIN 2007 "Efficient use of auxiliary information at the design and at the estimation stage of complex surveys: methodological aspects and applications for producing official statistics".

References

Crainiceanu C.M., Ruppert D., and Wand M.P. (2005). Bayesian analysis for penalized spline regression using WinBUGS. *Journal of Statistical Software*, **14**.

Lazarsfeld, P.F., and Henry, N.W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Malec, D., Sedransk, J., Moriarity, C.L. and LeClere, F.B. (1997). Small area inference for binary variables in the National Health Interview Survey. *Journal of the American Statistical Association*, **92**.

Montanari, G.E., Ranalli, M.G., and Eusebi, P. (2009). Latent variable modeling of disability in people aged 65 or more. *Manuscript*.

Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G., and Breidt, F.J. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society, Series B*, **70**.

The dynamic coregionalization model with application to air quality remote sensing

Alessandro Fassò[1], Francesco Finazzi[1]

[1] Dept. of IT and Mathematical Methods, University of Bergamo, Via Marconi 5, 24044 Dalmine BG, Italy. alessandro.fasso@unibg.it

Abstract: In this paper, we discuss the dynamic coregionalization model and its capability for model selection inference and interpretation in relation to spatio-temporal dynamic calibration and mapping of daily concentration of airborne particulate matter. To do this, we consider the problem of joint modelling ground level concentration data and satellite measurements of aerosol optical thickness (AOT), which are rarely available. The maximum likelihood estimation for the large data set related to the "padano-veneto" region, North Italy, with missing data is covered by the stable EM algorithm and implemented on a small size computer cluster.

Keywords: EM algorithm; maximum likelihood estimation; multivariate spatio-temporal missing data; particulate matters; aerosol optical thickness.

1 Introduction

The increasing availability of large datasets on multivariate spatio-temporal data parallels the need for statistical models which are flexible enough for covering the underlying complexity and can be estimated by means of well founded inferential techniques. The dynamic coregionalization model, recently proposed by Fassò et al. (2009), has these advantages as it allows modelling of complex multivariate spatio-temporal dynamics and performing maximum likelihood parameter estimation by means of the EM algorithm. Moreover, it naturally covers large amounts of "structural" missing data. This is particularly important for remote sensing applications where, under cloud conditions, the satellite data are missing.

2 Dynamic coregionalization model

We consider multivariate data which are cross-correlated at each point in geographical space, say D, and discrete time $t = 1, 2, ..., T$. Each variable is allowed to have a different spatial correlation and/or serial correlation over time. This is different to standard application of the coregionalization model to spatio-temporal data where it is commonly considered continuous

time, see e.g. De Iaco et al. (2005). In other words, we suppose that at time t, the observed data follow the equation

$$Y_t = X_t\beta + KZ_t + \bar{W}_t + \varepsilon_t \tag{1}$$

Ignoring for a while missing data, the observed Y_t is a $N-$dimensional vector containing the maps related to the q observed variables. Namely $Y_t = (Y_1(S_1,t),...,Y_q(S_q,t)) = (Y_i(s_{i,j},t)))'_{j=1,...,n_i,i=1,...,q}$ so that each variable Y_i is observed at sites $S_i = \{s_{i,1},...,s_{i,n_i}\}$ and $N = n_1 + ... + n_q$.
The first term of the RHS of equation (1), X, is given by a set of known covariates. The second term, Z, covers for the time dynamics being a stable multivariate Markov process in the form $Z_t = GZ_{t-1} + \eta_t$, $\eta_t \sim N(0,\Sigma_\eta)$. The Gaussian error ε is a white noise process with diagonal variance-covariance matrix Γ_0 whose elements are $\sigma^2_{\varepsilon,i}$, $i = 1,...,q$.
Finally, the third term of RHS of equation (1) is a zero mean $q-$dimensional Gaussian process $\bar{W}(s,t) = \left(\bar{W}_1,...,\bar{W}_q\right)$ defined by the so called linear coregionalization model with c components, namely $\bar{W}(s,t) = \sum_{p=1}^{c} W_p(s,t)$ where $W_p(s,t) = (W_{p,1},...,W_{p,q})$ is white noise in time but correlated over space with a $q \times q$ covariance and cross-covariance matrix function given by $\Gamma_p(h,\theta_p) = (\mathrm{cov}(W_{p,i}(s),W_{p,j}(s')))_{i,j=1,...,q} = V_p\rho_p(h,\theta_p)$ where $h = \|s - s'\|$ is the Euclidean distance. For each $p = 1,...,c$, V_p is a positive semi-definite $q \times q$ matrix and $\rho_p(h,\theta_p)$ is a valid correlation function, characterized by the parameter vector θ_p. In the sequel, the exponential correlation function is considered, namely $\rho_p(h,\theta_p) = \exp(-h/\theta_p)$. In addition, the processes W_p and $W_{p'}$ are uncorrelated so that the multivariate $q \times q$ covariance matrix for W is given by $\Gamma_{\bar{W}}(h,\theta_1,...,\theta_c) = \sum_{p=1}^{c}\Gamma_p(h,\theta_p) = \sum_{p=1}^{c} V_p\rho_p(h,\theta_p)$.
The model parameters are collected in the vector Ψ, which ignores duplications, namely $\Psi = vec^*(\beta,\Gamma_0;G,\Sigma_\eta,;V_1,\theta_1,...,V_c,\theta_c) = (\Psi_Y,\Psi_Z,\Psi_W)$.

3 Estimation and inference

Due to the Markovian assumption and to the space-time separability property of the model, the complete-data log-likelihood function takes the nice additive form

$$l(\Psi;Y,Z,W) = l(\Psi_Y;Y \mid Z,W) + l(\Psi_Z;Z) + l(\Psi_W;W) \tag{2}$$

However, (2) is not easily handled since Z is latent and Y is partially missing. The problem is overcome by considering the EM algorithm, already used by Fassò and Cameletti (2010) for univariate spatio-temporal environmental data and extended by Fassò et al. (2009) for the dynamic coregionalization model with missing data.
At the E-step of the algorithm, denoting by $Y^{(1)}$ the subset of actual observations, the expectation of the complete data log-likelihood under the parameter $\Psi^{(k)}$, conditionally to the observed data $Y^{(1)}$, is computed thanks to the iterated expectation theorem, that is

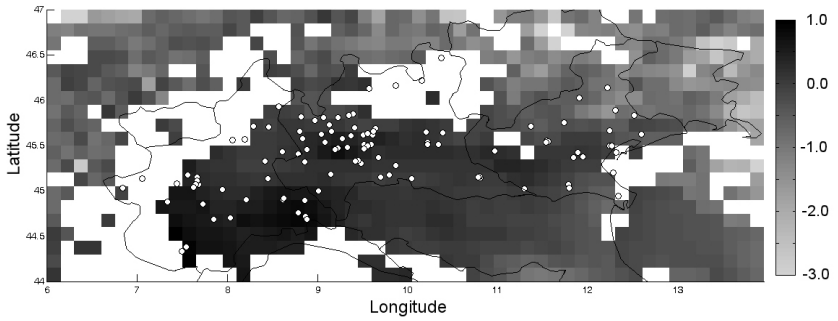


FIGURE 1. AOT standardized data (July 14th 2006) and ground level PM_{10} monitoring network sites (white circles).

$$Q\left(\Psi, \Psi^{(k)}\right) = E_{\Psi^{(k)}}\left[E_{\Psi^{(k)}}\left[l(\Psi; Y, Z, W) \mid Y, Z, W\right] \mid Y^{(1)}\right]$$

At the M-step, $Q\left(\Psi, \Psi^{(k)}\right)$ is maximized with respect to Ψ and $\Psi^{(k+1)}$ is chosen so that $\Psi^{(k+1)} = \arg\max Q\left(\Psi, \Psi^{(k)}\right)$. The solution of the maximization problem gives rise to quasi closed-form formulas for the update of Ψ, reported in detail in Fassò et al. (2009).
Since the EM algorithm is only guaranteed to converge to local maxima of the likelihood function, the whole estimation procedure is based on a set of EM estimation runs, each one characterized by different initial values for the parameter vector. Initial values are first evaluated through an estimation procedure based on the method of moments, detailed in Fassò and Finazzi (2010), and then locally perturbed by means of a random noise.
As the solution of the estimation procedure, the parameter vector $\hat{\Psi}$ is considered that gives rise to the maximum marginal log-likelihood $l_{Y^{(1)}}(\hat{\Psi})$. The role of the marginal log-likelihood, evaluated through the Kalman filter approach reported in Fassò et al. (2009), is important for computing likelihood-ratio tests and comparing nested models. Similarly, the estimated parameter vector $\hat{\Psi}$ is completed with standard deviations obtained by explicit recursive formulas for the Hessian matrix of the same marginal likelihood.

4 The case study

We consider ground-level data on concentration of airborne particulate matters PM_{10}, coming from $n_2 = 107$ monitoring stations. Although each station provides direct and reliable measures of the PM_{10} concentration, they have irregular spatial patterns. For this reason, a second variable is

model	M_0	M_1	M_2	M_3
$l_{Y^{(1)}}(\hat{\Psi})$	11271	21682	22543	22830
$Bias$	-0.0136	-0.0026	-0.0026	-0.0026
MSE	0.5004	0.2218	0.2214	0.2213

TABLE 1. Marginal log-likelihood and cross-validation results for models with $c = 0, ..., 3$ coregionalization components.

considered, namely the Aerosol Optical Thickness (AOT), which is known to be related with the particulate matters concentration and is useful to improve mapping capability of the PM_{10} concentration over the area of interest, see e.g. Koelemeijer et al. (2006).
AOT data are collected by the Terra and Aqua NASA satellites by means of the MODIS instrument (Moderate Resolution Imaging Spectroradiometer) and are provided with a spatial resolution of 10×10 km at nadir. The data set considered here covers the Italian region known as the padano-veneto area, bounded by a box of coordinates 44°N-6°E, 47°N-14°E, giving a daily data vector of 1134=54x21 elements, and the time period between March 2006 and September 2006 (see Figure 1). The daily average missing data rate for the AOT variable is 73% while it is 3% for the PM_{10}.
In order to improve calibration capability, several covariates are considered, including mixing height, accumulation of rain precipitation, land elevation, longitude of the site and percentage of urban area. PM_{10} concentrations and AOT measures are first log-transformed and then standardized, giving all variables with unit variance. Standardization is also applied to each covariate separately.

4.1 Model estimation and selection

In order to evaluate the role of the latent spatial variable W, models M_0, M_1, M_2 and M_3 are considered, with $c = 0, ..., 3$ coregionalization components respectively. It is worthwhile to note that, without coregionalization components, the spatial correlation function between sites is not directly modelled, though *lato sensu* a quota of the spatial correlation is covered by the covariates.
Models are estimated by means off the estimation procedure described in the previous section and compared by implementing likelihood-ratio tests between nested models. In order to evaluate the spatial prediction capability of each model, the leave-one-out crossvalidation method is applied over the PM_{10} sites S_2. Prediction bias and MSE are evaluated at each site $s_{2,j} \in S_2$. To do this, we estimate the model considering all data except the PM_{10} concentrations collected at $s_{2,j}$. The estimated model is then used to predict the PM_{10} concentration at $s_{2,j}$ for each day. Map average bias and map average MSE are reported in Table 1.

	$\hat{\beta}^{AOT}_{const}$	$\hat{\beta}^{AOT}_{MH}$	$\hat{\beta}^{AOT}_{Ele}$	$\hat{\beta}^{AOT}_{Urb}$	$\hat{\beta}^{AOT}_{Rain}$	$\hat{\beta}^{AOT}_{Long}$
value	−0.360	−0.143	−0.292	0.020	0.115	−0.005
std	0.140	0.009	0.006	0.002	0.010	0.001
	$\hat{\beta}^{PM}_{const}$	$\hat{\beta}^{PM}_{MH}$	$\hat{\beta}^{PM}_{Ele}$	$\hat{\beta}^{PM}_{Urb}$	$\hat{\beta}^{PM}_{Rain}$	$\hat{\beta}^{PM}_{Long}$
value	−0.097	−0.065	−0.133	0.106	−0.030	−0.133
std	0.136	0.010	0.006	0.004	0.011	0.005
	$\hat{\sigma}^2_{\varepsilon,AOT}$	$\hat{\sigma}^2_{\varepsilon,PM}$	$\hat{g}$	$\hat{\sigma}^2_{\eta}$		
value	0.041	0.192	0.880	0.084		
std	0.001	0.002	0.029	0.012		
	$\hat{v}^{AOT}_1$	$\hat{v}^{PM}_1$	$\hat{v}^{AOT,PM}_1$	$\hat{\theta}_1$		
value	0.923	0.367	0.177	162.194		
std	0.003	0.015	0.015	6.521		

TABLE 2. Estimated parameters and standard deviations for model M_1

4.2 Model interpretation

The map average MSEs of Table 1 suggest an improvement of the model predictive performance when the coregionalization variable $\bar{W}$ is considered. Indeed, the percentage of explained variance increases from 50 to 78 moving from M_0 to M_1. On the other hand, the performance is not significantly different when the number of coregionalization components is increased from one to either two or three.

If map prediction of the PM_{10} concentration is of concern, the more parsimonious M_1 model should be preferred, despite the likelihood-ratio test favouring, in this case, the unrestricted models with a near zero p-value.

Table 2 reports the estimated parameters along with their standard deviations. The $\hat{\beta}$ coefficients are directly comparable with each other since each covariate is standardized with respect to each variable. Note the opposite signs of $\hat{\beta}_{Rain}$, which is negative for PM, as precipitation usually reduces ground level concentrations, but is positive for AOT, due to the optical effect of those rare rainy days with non missing AOT data. Note also the difference in $\hat{\beta}_{Urb}$, which is related to the lower spatial resolution of AOT, so that single AOT pixels can include both urban and rural areas. Finally, the positive values of $\hat{\beta}$ suggest an east-west trend on the average PM_{10} concentration. In fact, the eastern side of the region considered is less urbanized and it is open to the winds from the Adriatic sea, while the western side is closed in by the Alps and is characterized by deficient air circulation. This aspect is confirmed by the positive sign of $\hat{g}$, related to the temporal dynamics. Net of the effect of covariates and time dynamics, the estimated cross-correlation between AOT and PM_{10}, based on matrix $\hat{V}_1$, is 0.30, which is consistent with marginal correlation of PM and AOT. The latent variable is also characterized by a persistent spatial correlation, described by the exponential correlation function with parameter $\hat{\theta}_1 \cong 162$

km.

5 Conclusions

We discussed the use of the dynamic coregionalization model in the framework of large dataset linear modelling for multivariate air quality data. Despite the large size of the data, the complex structure of the model and high rate of missing data, it is seen that this approach can be implemented with relatively standard computing facilities. Moreover, it covers in a natural way all inferential tools, including maximum likelihood estimation, classical likelihood inference, such as likelihood ratio tests, confidence intervals and crossvalidation. Of course, due to the large number of degrees of freedom, p-values have to be interpreted *cum grano salis*. Extensions to spatial nonstationarity and/or nonseparability can be naturally based on this model using the loess semiparametric approach of Bodnar and Schmid (2009) or the transformation based approach of Bruno et al. (2008).

References

Bruno F., Guttorp P., Sampson P.D., Cocchi D. (2008) A simple non-separable, non-stationary spatiotemporal model for ozone *Environmental and Ecological Statistics*. **16**, 515-529.

De Iaco, S., Palma, M., Posa, D. (2005). Modeling and prediction of multivariate space-time random fields. *Computational Statistics and Data Analysis,* **48**, 525-547.

Fassò, A., Cameletti, M. (2010). A unified statistical approach for simulation, modelling, analysis and mapping of environmental data. *Simulation*. **86**, 3, 139154.

Fassò, A., Finazzi, F. (2010). Statistical mapping of air quality by remote sensing: uncertainty and sensitivity to missing data. Submitted.

Fassò, A., Finazzi, F., D'Ariano, C. (2009) . Integrating satellite and ground level data for air quality monitoring and dynamical mapping. GRASPA Working Paper n.34. (www.graspa.org).

Koelemeijer, R.B.A., Homan, C.D., Matthijsen, J. (2006). Comparison of spatial and temporal variations of aerosol optical thickness and particulate matter over Europe. *Atmospheric Environment*, **40**.

Bodnar, O., Schmid W. (2009). Nonlinear locally weighted kriging prediction for spatio-temporal environmental processes *Environmetrics*. DOI: 10.1002/env.1005

Improving estimative prediction regions

Giovanni Fonseca[1], Federica Giummolè[2], Paolo Vidoni[1]

[1] University of Udine, Department of Statistics, via Treppo 18, I-33100 Udine, ITALY
[2] University of Venice, Department of Statistics, San Giobbe, Cannaregio 783, I-30121 Venice, ITALY

Abstract: In this work we address the problem of prediction in a multidimensional setting. Generalizing a result presented in Ueki and Fueda (2007), we propose a method for correcting estimative predictive regions to reduce their coverage error to third order accuracy. The improved prediction regions are easy to calculate using a suitable bootstrap procedure. An example of application is also included, showing the performance of the proposed method.

Keywords: Coverage probability, Parametric bootstrap, Prediction regions.

1 Introduction

Let $Y = (Y_1, \ldots, Y_n)$ be an observable random vector. The problem of prediction consists on giving an estimate of a further random variable Z, on the basis of an observed sample y from Y. The joint distribution of Z and Y is assumed to be known, up to a k-dimensional parameter $\omega \in \Omega \subset \mathcal{R}^k$. In the following, $\hat{\omega} = \hat{\omega}(Y)$ denotes the maximum likelihood estimator (MLE) for ω. In the case of Z being a unidimensional continuous random variable, a possible solution can be given in terms of prediction limits, i.e. functions $\tilde{z}_\alpha(\hat{\omega})$ such that

$$P_{Y,Z}\left[Z \leq \tilde{z}_\alpha(\hat{\omega}(Y))\right] = \alpha,$$

for all $\alpha \in (0,1)$, at least to a high order of approximation. The above probability is intended with respect to the joint distribution of Z and Y and is called coverage probability. Exact results are usually linked to the existence of a pivotal quantity and can only be found in special cases. An easy solution is given by considering the estimative prediction limits, obtained by substituting the unknown parameter ω by $\hat{\omega}$ in the α-quantiles of the conditional distribution of Z given $Y = y$. Unfortunately the associated coverage error has order n^{-1}, which is often considerable. Prediction limits with coverage error of order $o(n^{-1})$ have been proposed in Barndorff-Nielsen and Cox(1996) and Vidoni (1998), as modifications of the estimative prediction limits. Recently, Ueki and Fueda (2007) suggested a simulation-based procedure, useful to easily compute improved prediction limits. In this work

we extend their result to the case of Z being an multidimensional random variable. We define improved prediction regions as modifications of estimative prediction regions, where the correction term is easy to compute using a suitable parametric bootstrap simulation.

2 Improved prediction regions

Let us assume, for simplicity, that $Y_1, \ldots, Y_n$ and $Z = (Z_1, \ldots, Z_m)$ are independent random vectors. We denote by $f(z;\omega)$ and $F(z;\omega)$ the joint density and distribution function of Z, respectively. The problem of prediction in a multidimensional setting can be addressed through prediction regions, that are suitable subsets of $\mathcal{R}^m$ with a fixed probability of including Z. In particular, as suggested in Ueki and Fueda (2007), we can consider predictive regions of the form

$$D(r, \hat{\omega}) = \{z \in \mathcal{R}^m : R(z, \hat{\omega}) \leq r\},$$

for some real value r and some smooth real function $R(z, \omega)$. Notice that the so-called highest prediction density (HPD) regions, obtained by profiling the estimative predictive density $f(z;\hat{\omega})$, are a special case with $R(z, \omega) = -f(z;\omega)$. Once chosen the function $R(z, \omega)$, every prediction region of this form is identified by the value of r, which we refer to as the limit of the region. From now on, we assume that $R(z, \omega)$ is given and we focus our attention on finding a suitable value for r. Thus, our aim is to find a prediction limit, that is a function $\tilde{r}_\alpha(\hat{\omega})$ such that

$$P_{Y,Z}\left[R(Z, \hat{\omega}(Y)) \leq \tilde{r}_\alpha(\hat{\omega}(Y))\right] = E_Y\left[\int_{D(\tilde{r}_\alpha(\hat{\omega}),\hat{\omega})} f(z;\omega)dz\right] = \alpha + o(n^{-1}).$$

The above probability is the coverage probability of the prediction region associated with the limit $\tilde{r}_\alpha(\hat{\omega})$. Notice that, when Z is unidimensional and $R(Z, \omega) = Z$, $\tilde{r}_\alpha(\hat{\omega})$ corresponds to the usual α-prediction limit for Z, considered in Barndorff-Nielsen and Cox (1996) and Vidoni (1998).
As in the unidimensional case, an easy solution consists on working with the estimative prediction density $f(z;\hat{\omega})$. Let $\hat{r}_\alpha(\hat{\omega})$ be the estimative prediction limit, such that

$$\int_{D(\hat{r}_\alpha(\hat{\omega}),\hat{\omega})} f(z;\hat{\omega})dz = \alpha,$$

and let $\hat{\alpha}(\omega)$ denote the coverage probability associated to the corresponding estimative prediction region:

$$\hat{\alpha}(\omega) = E_Y\left[\int_{D(\hat{r}_\alpha(\hat{\omega}),\hat{\omega})} f(z;\omega)dz\right].$$

Unfortunately, $\hat{\alpha}(\omega) = \alpha + O(n^{-1})$. In order to eliminate the $O(n^{-1})$ term, we need to modify the estimative limit $\hat{r}_\alpha(\hat{\omega})$, considering instead a prediction limit of the form $\tilde{r}_\alpha(\hat{\omega}) = \hat{r}_\alpha(\hat{\omega}) + d(\hat{\omega})/n$, for a suitable function $d(\omega)$. Following Ueki and Fueda (2007), we propose to calculate the correction term as $d(\hat{\omega})/n = \hat{r}_\alpha(\hat{\omega}) - \hat{r}_{\hat{\alpha}(\omega)}(\hat{\omega})$. In fact, it can be shown that this correction term leads to a prediction limit with associated coverage error of order $o(n^{-1})$. Thus, the adjusted prediction limit improves on the estimative one, as far as coverage probability is considered. The improved prediction limit can be finally written as

$$\tilde{r}_\alpha(\hat{\omega}) = 2\hat{r}_\alpha(\hat{\omega}) - \hat{r}_{\hat{\alpha}(\omega)}(\hat{\omega}).$$

In order to explicitely calculate $\tilde{r}_\alpha(\hat{\omega})$, we only need to evaluate the estimative coverage probability $\hat{\alpha}(\omega)$. This can be easily computed in practice, using a suitable parametric bootstrap procedure.

3 Example

Let $Y_1, \ldots, Y_n$ and Z be independent random vectors with the same multivariate normal distribution $\mathcal{N}_p(\underline{0}, \Sigma)$, with Σ a symmetric non singular unknown variance matrix. We denote by Y the $n \times p$ matrix which i-th row is Y_i^T. Here, ^{T} means transpose and vectors are considered as column vectors. We denote by $\hat{\Sigma} = Y^T Y/n$ the sample variance.
Notice that in this situation we are able to find an exact solution. Indeed, $\frac{n-p+1}{np} Z^T \hat{\Sigma}^{-1} Z$ is a pivotal quantity, having Fisher F distribution $F(p, n-p+1)$ (see Mardia et al., 1979, chapter 3). Thus a region with exact coverage probability α can be written as

$$\left\{ z \in \mathcal{R}^p : z^T \hat{\Sigma}^{-1} z \leq \frac{np}{n-p+1} f_{\alpha,p,n-p+1} \right\},$$

where $f_{\alpha,p,n-p+1}$ is the α-quantile of a $F(p, n-p+1)$ distribution. Nonetheless, the aim of this example is to test the performance of the proposed method by comparing the coverage probabilities associated to improved prediction regions with those relative to estimative ones. In order to do that, we consider prediction regions obtained by profiling the estimative predictive density. For the normal distribution, these can be written in the form

$$D(r, \hat{\Sigma}) = \{ z \in \mathcal{R}^p : z^T \hat{\Sigma}^{-1} z \leq r \}.$$

It is a known result that $Z^T \Sigma^{-1} Z$ has a Chi-squared distribution χ^2_p. Thus, the estimative index $\hat{r}_\alpha(\hat{\Sigma})$ coincides with the α-quantile of a χ^2_p distribution, $\chi^2_{\alpha,p}$. The corresponding coverage probability, $\hat{\alpha}(\Sigma) = P_{Y,Z}(Z^T \hat{\Sigma}^{-1} Z \leq \chi^2_{\alpha,p})$, can be evaluated by means of a suitable parametric bootstrap procedure. Now, the improved prediction limit can be easily calculated as

$$\tilde{r}_\alpha(\hat{\Sigma}) = 2\hat{r}_\alpha(\hat{\Sigma}) - \hat{r}_{\hat{\alpha}(\Sigma)}(\hat{\Sigma}) = 2\chi^2_{\alpha,p} - \chi^2_{\hat{\alpha}(\Sigma),p}.$$

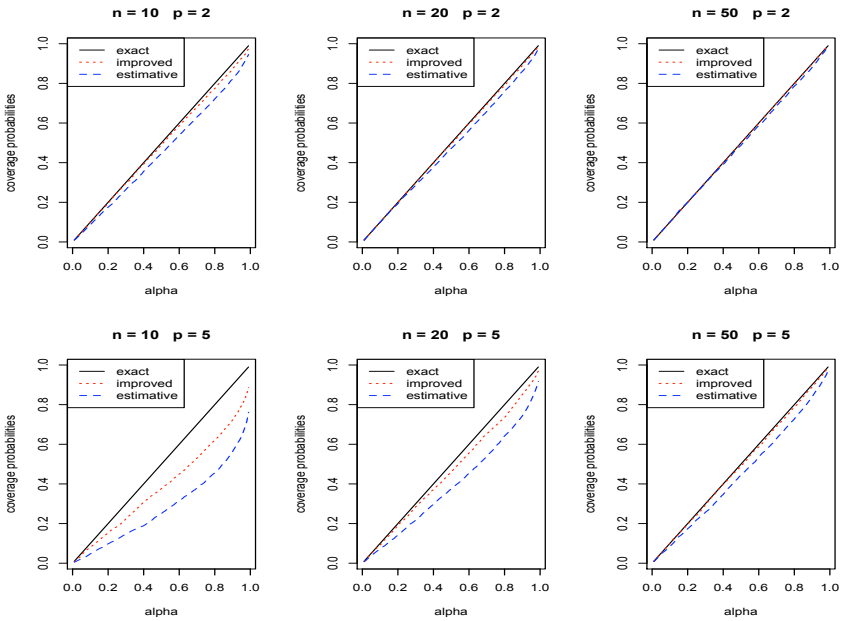


FIGURE 1. Multinormal prediction. Coverage probability plots for estimative (dashed) and improved (dotted) prediction regions, together with the diagonal line (solid), for different values of the sample size n and the dimension p of the multivariate normal distribution.

In a simulation study, the coverage probabilities of estimative and improved prediction regions have been estimated, for different values of α, using 5000 bootstrap replications. In Figure 1, the results show the superiority of the proposed adjusted prediction regions on the estimative ones.

References

Barndorff-Nielsen, O.E., and Cox, D.R. (1996). Prediction and asymptotics. *Bernoulli*, **2**, pp. 319-340.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979). *Multivariate Analysis*. London: Academic Press.

Ueki, M., and Fueda, K. (2007). Adjusting estimative prediction limits. *Biometrika*, **94**, pp. 509-511.

Vidoni, P. (1998). A note on modified estimative prediction limits and distributions. *Biometrika*, **85**, pp. 949-953.

Modelling the evolution of the number of armed conflicts

Sara Fontdecaba[1], Pere Puig[2], Jordi Valero[1]

[1] Universitat Politècnica de Catalunya, Avd. Diagonal 647, 08028 Barcelona, Spain. Email: sara.fontdecaba@upc.es
[2] Universitat Autònoma de Barcelona, 08193, Cerdanyola del Vallès (Barcelona), Spain.

Abstract: The discrete time series of the number of active conflicts by year, with at least 1000 battle related deaths, is studied by using an INAR(1) model with Hermite innovations. The parameters are estimated by using a robust moment based method and maximum likelihood as well. Several problems of forecasting are also considered in this work.

Keywords: INAR model; Hermite distribution; discrete time series; binomial thinning; overdispersion; forecasting

1 Motivation

The use of statistical models based on time series of counts has been increasingly popularized in many fields. Traditionally the count data with time structure has been modelled adapting the well known autoregressive AR processes to the integer-valued case, trough binomial thinning operator, leading to the INAR models. Commonly, these models consider the assumption of having a Poisson marginal distribution. However, overdispersion is frequent in practice and this assumption could be rejected. We propose a more general INAR(1) model where the marginals are Hermite distributed and, consequently, it includes the classical INAR(1) Poisson based model.
The analyzed example corresponds to the stationary yearly time series of the number of active conflicts. The studied conflicts are those that caused at least 1000 battle related deaths, in some sense the minimum number of victims to be considered as a war. The data set, from 1946 to 2008, gives an idea of the evolution of extrasystemic, interstate, intrastate and internationalized global conflicts. We calculate the maximum likelihood estimators of the parameters of this generalized INAR(1) model and interpret the results. These are also compared with those of the robust moment based method.

2 The Model

2.1 Model definition

The process X_t is called an INAR(1) model if it follows the recursion

$$X_t = p \circ X_{t-1} + W_t, \tag{1}$$

where p is a fixed parameter $p \in [0, 1)$ and W_t is an independently and identically distributed (*iid*) sequence of discrete random variables. W_t and X_{t-1} are presumed to be independent for all points in time. The discreteness of the process X_t is guarantied by the *binomial thinning* operation (Steutel and van Harn, 1979):

$$p \circ X_{t-1} \equiv Y_{1,t-1} + Y_{2,t-1} + \cdots + Y_{X_{t-1},t-1} = \sum_{i=1}^{X_{t-1}} Y_{i,t-1} \tag{2}$$

where $Y_{i,t-1}$ are assumed to be iid Bernoulli random variables with $P(Y_{i,t-1} = 1) = p$ and $P(Y_{i,t-1} = 0) = 1 - p$. Since $p \circ X_{t-1}$ given X_{t-1} is a sum of i.i.d Bernoulli random variables, it follows a Binomial distribution with parameters X_{t-1} and p. These processes have been considered extensively elsewhere (for instance, see McKenzie, 2000 and Brannas and Hellstrom, 2001).

In our case study we assume that the innovations (W_t) follow a Hermite distribution and it implies that the marginals are Hermite distributed as well. It is useful because in practice the variance is usually larger than the mean, that is, there is overdispersion. The Hermite distribution was introduced by Kemp and Kemp, (1966) and this is a two-parameter family that can be parameterized by the mean μ and the coefficient of dispersion δ. It will be denoted as $H(\mu, \delta)$.

2.2 Parameter estimation

In order to find the maximum likelihood estimates of the parameters for this INAR(1) model with Hermite innovations, we take into account the first order dependence of the process. It leads to the following simplified likelihood function:

$$L(X; \theta) = f(x_1, x_2, \cdots, x_n) = f(x_1) f(x_2|x_1) f(x_3|x_2) \cdots f(x_n|x_{n-1}) \tag{3}$$

Consequently, it is necessary to calculate the probability function of $X_i|X_{i-1}$, that is, the probability function of the sum of a Binomial(X_{i-1}) and a $H(\mu, \delta)$ both independent. Therefore, the likelihood function can be expressed as:

$$L(X; \mu, \delta) = P(X_1) \prod_{i=2}^{n} P(X_i|X_{i-1}) \tag{4}$$

assuming that the first observation follows the marginal or limit distribution. It can be shown, trough the properties of the probability generating function, that this is a $H(\frac{\mu}{1-p}, \frac{p+\delta}{1+p})$.

2.3 Forecasting

One of the most commons procedures for forecasting the mean value is to use conditional expectations. Applying the properties of the *binomial thinning operator* one can show that the distribution of $X_{t+h}|X_{t-1}$ is the sum of a binomial distribution with parameters p^{h+1} and X_{t-1} and a Hermite distribution of mean $\sum_{h=0}^{h} p^h \mu$, both independent. Consequently, this result leads to the following predictor of the mean:

$$\widehat{X}_{t+h}|X_{t-1} = E[X_{t+h}|X_{t-1}] = p^{h+1}X_{t-1} + \sum_{h=0}^{h} p^h \mu, \quad h = 0, 1, 2, ...,$$

Note that as $h \to \infty$, this predictor reaches the value $\mu/(1-p)$, the marginal mean of the process. In practice the values of μ and p will be replaced by their maximum likelihood estimates $\hat{\mu}$ and $\hat{p}$, and the delta method will be used in order to calculate the variance of the predictions.

3 Analysis of the Armed Conflicts

This data set corresponds to the counts of the global active armed conflicts (source UCDP: Uppsala Conflict Data Program. Gleditsch, 2002) with at least 1000 battle-related deaths in a given year. UCDP defines conflict as a contested incompatibility that concerns government and/or territory where the use of armed force between two parties, of which at least one is the government of a state, results in at least 25 battle-related deaths. The presented study during the period 1948-2008, only includes those that the intensity is high, and more than 1000 victims have been registered. The original database also includes information of the two priors years, 1946 and 1947, dismissed in this study as possible consequences of The World War II.

The absence of a clear time pattern and the lack of an increasing trend over time, indicate that the series seems to be stationary. During these 61 years, the behavior of the number of conflicts is consistent with an average of 9 conflicts per year without any alteration to remark.

In this example, the structure of the INAR(1) recursion can be interpreted as the number of conflicts at time t, obtained by means of the number of conflicts that remain active at $t-1$ plus the new conflicts that appear at the same time t. In this sense, the parameter p can be interpreted as the percentage of conflicts that remain active in two consecutive years. The parameter μ gives an idea of the average number of new conflicts that begin each year and parameter δ indicates the coefficient of dispersion of these new conflicts.

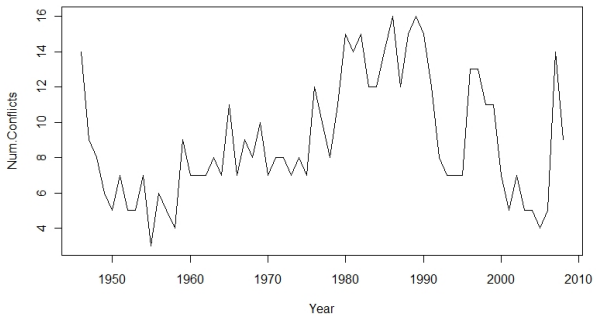


FIGURE 1. Number of armed conflicts during the period 1948-2008

3.1 Estimation Results

We first check that a first autoregressive order is enough to explain the correlation pattern, trough the shape of the ACF (Auto Correlation Function) and the PACF (Partial Auto Correlation Function). Afterwards we estimate the parameters by using the robust moment based method. The results have been the following: $\tilde{p} = 0.7215$, $\tilde{\mu} = 2.4789$ and $\tilde{\delta} = 1.5715$. Note that the estimate of the dispersion index is greater than 1, the theoretical value under the Poisson assumption for the innovations. Consequently, it is reasonable to use the $H(\mu, \delta)$ distribution for these innovations.
Therefore, we maximize (4) with a program developed in R that is available from the authors upon request. The starting points for the numerical procedure are the values coming from the moment based method. The obtained estimates, with their standard errors in brackets, has been the following: $\hat{p} = 0.7644$ (0.0540), $\hat{\mu} = 2.1091$ (0.5116) and $\hat{\delta} = 1.7426$ (0.4106). Note the improvement of the Hermite distribution ($\delta \in (1, 2)$) for W_t, in front of the Poisson distribution ($\delta = 1$). The one-tailed Wald test rejects the null hypothesis that innovations follows a Poisson distribution with $p = 0.035$.

3.2 Forecasting

Table 1 shows the predictions for 20 steps ahead jointly with their variance. The value for $h = \infty$ corresponds to $\hat{\mu}/(1 - \hat{p})$, the estimate of the mean of the process.
In order to validate the model, predicted and observed valued are compared for each time t. Figure 2 shows these predicted values (blue line) with their 95% confidence intervals. The 20 predictions in the future are also shown

TABLE 1. Mean predictions of the model and its variance

h	$\hat{X}_{t+h}$	$\hat{V}(\hat{X}_{t+h})$	h	$\hat{X}_{t+h}$	$\hat{V}(\hat{X}_{t+h})$
1	6.643	0.483	11	8.796	1.302
2	7.187	0.650	12	8.833	1.325
3	7.603	0.789	13	8.861	1.346
4	7.923	0.904	14	8.883	1.362
5	8.164	0.998	15	8.899	1.375
6	8.350	1.075	16	8.912	1.385
7	8.492	1.140	17	8.922	1.393
8	8.601	1.193	18	8.930	1.400
9	8.684	1.237	20	8.939	1.409
			∞	8.953	1.424

(red line). Note how these predictions tend to the estimated mean of the process (green line).

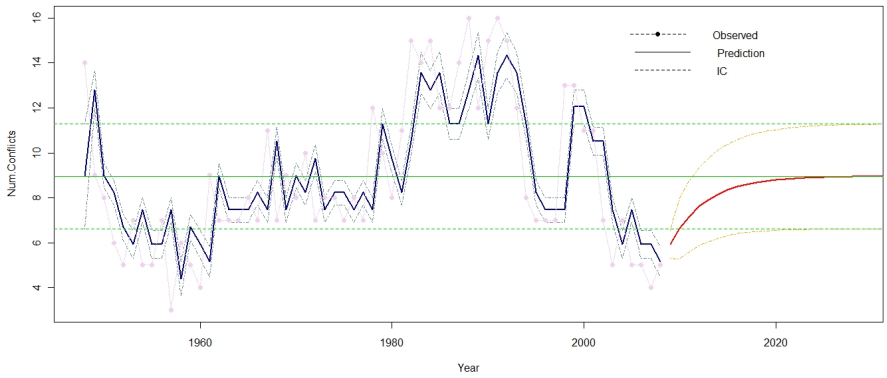


FIGURE 2. Prediciton of the mean of the process with its confidence interval

4 Concluding Remarks

We have presented a INAR(1) model with Hermite innovations that is useful to analyze the time series of the number of armed conflicts during the period 1946-2008. According to the obtained results, a 76% of the conflicts seem to remain active during the next year, being able to be related with non finished conflicts at the end of the current year. In average, there

are two new conflicts that have to be added to those that remain active from the previous year. The value of the dispersion parameter greater than one, obtained from the based moment estimation method, indicates that Hermite distribution can be used to fit this data set. According to this model the number of expected conflicts for 2010 would be between 6 and 9.

Acknowledgments: This work was supported by the Comissionat per a Universitats i Recerca del Departament d'Innovació, Universitats i Empresa de la Generalitat de Catalunya i del Fons Social Europeu and by grant MTM2009-10893 from the Ministry of Education of Spain.

References

Brännäs, K., Hellström, J. (2001) *Generalized Integer-Valued Autoregression* Econometric Reviews, Vol. 20, Number 4, pp. 425-443. ed Publisher Taylor & Francis.

Gleditsch, N. P. et al. (2002). *Armed Conflict 19462001: A New Dataset.* Journal of Peace Research, Vol. 39, Number 5, pp. 615-637.

Kemp, C.D., Kemp, A.W. (1965).*Some properties of the 'Hermite' Distribution.* Biometrika, Vol. 52, Number 3-4, pp. 381-394.

McKenzie, E. (2000).*Discrete variate time series.* Handbook of Statistics, Vol. 21, pp. 573-606.

Steutel, F.W and Van Harn, K. (1979).*Discrete analogues of self-decomposability and stability.* The Annals of Probability, Vol. 7, pp. 893-899.

Assessing the variability of Scottish rivers using wavelet analysis

Maria Franco-Villoria[1], Marian Scott[1], Trevor Hoey[2], Denis Smith[3]

[1] Department of Statistics, 15 University Gardens, University of Glasgow, G12 8QQ. Contact: mvilloria@stats.gla.ac.uk
[2] Department of Geographical and Earth Sciences, University of Glasgow
[3] Department of Management, University of Glasgow

Abstract: Wavelet analysis is presented here as a possible method for dealing with non-stationary environmental time series. The results obtained from a series of monthly maxima flows on the river Tweed indicate that even though there is a yearly cycle, it is not constant over the whole time period.

Keywords: stationarity; variability; flood; wavelet power spectrum

1 Introduction

1.1 Background

River flow records have formed the basis of many flood risk estimates, based on classical statistical models, that have assumed stationarity. However, under climate change and climate change scenarios, there is an expectation that the flow series may no longer be stationary, therefore statistical models that do not make this assumption are required. In this paper, wavelet modelling is applied to river flow series from Scotland.

1.2 Data set

The River Tweed is situated in the South East of Scotland and has a catchment area of 4390km^2. Data (gauged daily flow) were collected at Norham (Station 21009). In total, 16821 observations were available (1/10/62 - 31/10/08). The time series was log transformed to stabilize the variance. Since the main interest lies in the extreme values the series of monthly maxima was calculated. Monthly rainfall data (mm across catchment) was also available for 1961-2007. Data were provided by the National River Flow Archive and the Scottish Environment Protection Agency (SEPA).

2 Methods - Wavelet Analysis

One way of identifying the local behaviour of non-stationary time series is by wavelet analysis. By subsequently filtering the original series, we obtain sequences of results which relate to variations at different scales (frequencies). The result is a time-frequency representation of the data (Percival & Walden(2006)). The continuous wavelet transform of a time series $\{x_t\}$ at a particular localized time n and scale s is defined as:

$$W_n(s) = \sum_{t=0}^{N-1} x_t \psi * \left(\frac{(t-n)\delta t}{s} \right) \tag{1}$$

where * indicates the complex conjugate, δt is the time spacing of the series $\{x_t\}$ and $\psi(\cdot)$ is a scaled and translated version of the chosen wavelet function $\psi_o(\eta)$, which has been normalized previously. A key point is that $W_n(s)$ preserves all the information in the original time series x_t. A common example of a continuous wavelet function is the Morlet wavelet $\psi_0(\eta) = \pi^{-1/4} e^{i\omega_0\eta} e^{-\eta^2/2}$, where η is a nondimensional time parameter and ω_0 is a nondimensional frequency (ω_0=6 for the Morlet wavelet). An interesting application of the continuous wavelet analysis is the wavelet power spectrum (WPS), defined as $|W_n(s)|^2$. It provides a measure of the variability of the time series at each scale s and time t. Significance testing is available so that both particular scales and time periods with significant variability can be identified (Torrence & Compo(1998)). The scale averaged wavelet power over a range of scales s_{j_1} to s_{j_2} is defined as:

$$\overline{W}_n^2 = \frac{\delta j \delta t}{C_\delta} \sum_{j=j_1}^{j_2} \frac{|W_n(s)|^2}{s_j} \tag{2}$$

where δj is the scale spacing and C_δ is a reconstruction constant (C_δ=0.776 for the Morlet wavelet). The resulting series is a time series of the average variance in a certain frequency band that can be used to examine the variability at any particular set of scales (Torrence & Compo(1998)).

3 Results

3.1 River Tweed Analysis

Continuous wavelet analysis using the Morlet wavelet was applied to the monthly maxima river flow series (Figure 1) and rainfall series (Figure 2) to obtain the wavelet power spectra. These showed that the variability was concentrated on the 1 year band. However, the river flow spectrum was not constant across the whole time period, with periods of high (significant) variability (1976-81, 1988-98) alternated with periods of non-significant

variability. The rainfall spectrum showed a period of significant variability between 1993 and 1997, and patches of high (though not significant) variability both on the 1yr band and at lower scales. This is a clear indication of non-stationarity. The scale averaged wavelet power for scales 0.86-1.16yr was calculated for both the monthly maxima and the rainfall series to get a measure of fluctuations in variability in the 1yr band. The 1yr variability series are plotted on Figure 3. The red dots represent the most significant historical floods for this river before 1995 (Acreman (1989), Fox and Johnson (1997)). Flooding may arise under very different conditions and hence it is difficult to directly link the floods with particular features of the variability series on Figure 3. If we look at the two highest recorded floods (1982 and 1992)(Fox and Johnson (1997)) we can see that the 1982 flood coincides with a relatively high variability period for the rainfall series but a low-variability period for the flow series, while for the 1992 flood the variability is about the same for both series. Such a difference between the two floods is likely to be related to the conditions under which they were created (the 1982 event is associated with snowmelt (Fox and Johnson (1997)). The 1994 flood coincides with the highest peak in both flow and rainfall variability series; however, the difference between the two series variability is noticeable. This particular event was associated with intense rainfall over 48hours in the south of Scotland, affecting all major catchments in the region (Black and Bennett (1994)). The 1987 flood, which occurred during a period of low variability, is associated with high flows in the headwaters. The fact that our recording site is close to the mouth of the river would explain the low observed variability.

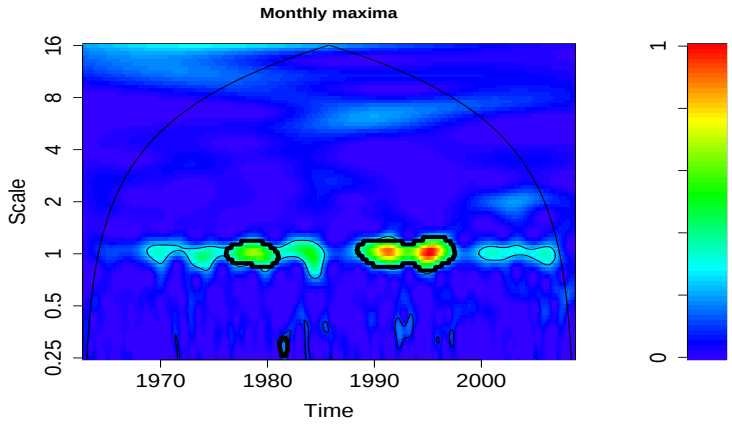


FIGURE 1. Wavelet power spectrum of monthly maxima series

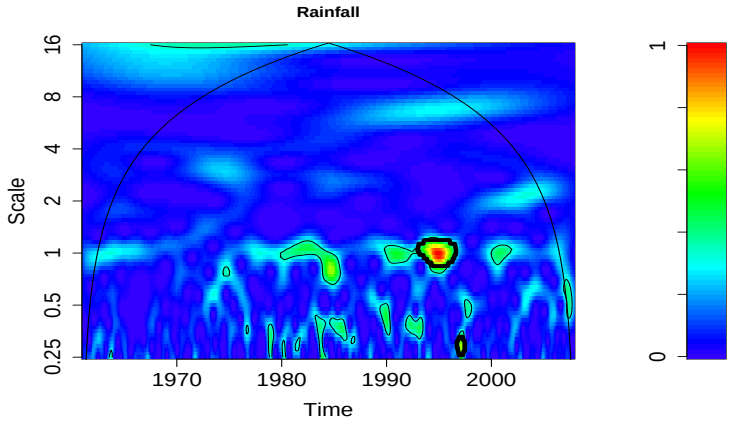


FIGURE 2. Wavelet power spectrum of monthly rainfall series

Looking at Figure 3, there is a clear change point around mid 1986, when the variability is minimum (followed by another period of very low variability around the beginning of 1998). This means that both rainfall and river flow were quite stable at that time. Grew & Werritty (1995) defined flood poor (1964-1973) and rich (late 1980s - early 1990s) periods. In particular, on the river Tweed, the 1970s were identified as being relatively dry, followed by a significantly wetter period in the late 1980s until the 1990s (Fox and Johnson (1997)), which coincide with low and high variability periods respectively on the graph. The period of time from 1977 to 1986 has been characterized as the wettest period on record for the UK as a whole (Marsh (1995)), which corresponds with the 'cluster' of high variability and flood events just before the low variability change point, followed by a second 'cluster' of peaks which corresponds to the flood-rich period. The decrease in variability since summer 1995 until it reaches a minimum point in 1998 would be explained by a remarkably dry period in April-August 1995 followed by a dry winter 1995/1996 which resulted in the second driest summer for Scotland (Marsh (1995)).

3.2 Summary and future work

Wavelet analysis is presented here as a powerful tool for exploring the variability of a non-stationary time series. The results not only agree well with previous studies and historical flood events but also provide a useful insight into the yearly variability of the series. The North Atlantic Oscillation and the Atlantic Meridional Oscillation will be explored as potential drivers of this variability. The methods presented here are being applied to different rivers across Scotland to investigate spatial and catchment scale differences.

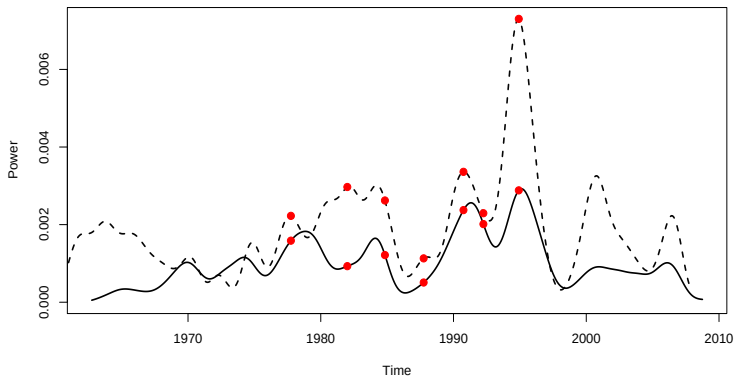


FIGURE 3. 1yr band variability series. Solid line represents flow, dashed line represents rainfall. Each flood is represented twice, once on each series.

References

Acreman, M.C. (1989). Extreme historical UK floods and maximum flood estimation. *Water and Environment Journal*, **3(4)**, 404-412.

Black, A.R. and Bennett, A.M. (1994). Regional flooding in Strathclyde December 1994. *Hydrological Data - Institute of Hydrology - Clyde River Purification Board*, 29-34.

Fox, I.A. and Johnson, R.C. (1997). The hydrology of the River Tweed. *The Science of the Total Environment*, **194-195**, 163-172.

Grew, H. and Werritty, A. (1995). Changes in flood frequency and magnitude in Scotland 1964-1992. In: *Proceedings of the BHS Fifth National Hydrological Symposium*. 3.1-3.9, Edinburgh.

Marsh, T.J. (1995). The 1995 drought - a water resources review in the context of the recent hydrological instability. *Hydrological Data UK Series. Institute of Hydrology, Wallingford*, 25-33.

Percival, D.B. and Walden, A.T. (2006). *Wavelet Methods for Time Series Analysis*. Cambridge University Press, Cambridge Series in Statistical and Probabilistic Mathematics.

Torrence, C. and Compo, G.P. (1998). A Practical Guide to Wavelet Analysis. *Bulletin of the American Meteorological Society*, **79**, 61-78.

Filtering and smoothing algorithms for discrete-time systems with multiple packet dropouts using covariance information

M. J. García-Ligero[1], A. Hermoso-Carazo[1] and J. Linares-Pérez[1]

[1] Dpto. de Estadística e I.O., Universidad de Granada, Avda Fuentenueva s/n, 18071 Granada, Spain. (mjgarcia, ahermoso, jlinares)@ugr.es

Abstract: The estimation problem of the signal based on measurements with multiple packet dropout is considered when the data arrival probability to a processing unit is known. Assuming that the equation which describes the state is unknown, we derive recursive algorithms for the filtering and smoothing problem using the information provided by the covariance functions of the processes involved in the measurement equation.

Keywords: Packet dropouts; Covariance information; Least-squares estimation.

1 Introduction

Recently, the trend of utilizing networks for transmitting measurement data have gained interest over the traditional systems of communication. This is due to the advantages presenting this transmission form such as low cost and flexibility, between others. However, in many practical situations, its use can carry about some problems such as lost data or packet dropouts due to unreliable characteristic of networks. Packet dropouts are a kind of uncertainty which can occurs randomly due to failures in transmission or network congestion. The packet dropouts can be modeled by independent Bernoulli random variables; under this assumption Sahebsera et al. (2007) study the estimation problem of the signal transforming the original system to a stochastic parameter system by augmentation of the state and measurement. Using this model, Sun et al. (2008) derive the optimal linear filter and smoother in the minimum variance sense by means an innovation approach.

The aforementioned papers study the estimation problem in multiple packet dropouts systems considering a full knowledge of state-space model. However, in many practical situations, the equation which describes the state is unknown and the estimation problem can be addressed using covariance information (see for example, Nakamori et al. (2006)). In this paper, we study the linear estimation problem of the signal based on measurements

with multiple packet dropouts using the information provided by the covariance functions of the processes involved in observation equation. The packet dropouts are modeled by independent Bernoulli random variables with known distribution. Under these assumptions, we derive recursive algorithms to obtain the filter and fixed-point smoother using that the covariance function of the signal is expressed in a semi-degenerate kernel form. These algorithms are obtained by using an innovation approach which, as it is known, provides a simple derivation of the estimation algorithms.

2 Problem formulation

Consider the measurement of a $n \times 1$ signal, x_k, described by the equation

$$z_k = x_k + v_k, \quad k \geq 1$$

where $\{v_k,\ k \geq 1\}$ is a white noise. We assume that the measurement, z_k, is transmitted to a processing unit through an unreliable network, where it is possible that some data are lost during the transmission. According to Sahebsara et al. (2007), this situation can be modeled as

$$y_k = \xi_k z_k + (1 - \xi_k) y_{k-1}, \quad k > 1; \quad y_1 = \xi_1 z_1 \tag{1}$$

where $\{\xi_k,\ k \geq 1\}$ is a sequence of independent Bernoulli random variables with $P[\xi_k = 1] = p_k$. If $\xi_k = 1$, the measurement at time k is received with probability p_k, whereas $\xi_k = 0$, which occurs with probability $1 - p_k$, the measurement at time k is lost and then the received observation is y_{k-1}. The following hypotheses are set on the signal and noise processes:

(I) The signal, $\{x_k,\ k \geq 1\}$, has zero mean and its covariance function is

$$E[x_k x_s^T] = \alpha_k \beta_s^T,\ s \leq k$$

where α_k and β_s are known $n \times m$ matrix functions.

(II) $\{v_k,\ k \geq 1\}$ is a white noise with zero mean and $E[v_k v_k^T] = R_k$.

(III) $\{\xi_k,\ k \geq 1\}$ is a sequence of independent Bernoulli random variables with known probabilities, $P[\xi(k) = 1] = p_k > 0,\ \forall k \geq 1$.

(IV) $\{x_k,\ k \geq 1\}$, $\{v_k,\ k \geq 1\}$ and $\{\xi_k,\ k \geq 1\}$ are mutually independent.

Our aim is to determine the least-squares linear estimator of the signal, x_k, based on the information provided by the measurements $\{y_1, \ldots, y_L\}$, given by (1); more specifically, the filter ($L = k$) and the fixed-point smoother (k fixed and $L > k$). The linear estimation problem is treated by using an innovation approach, which is based on the one-to-one correspondence between the observations, $\{y_1, \ldots, y_L\}$, and the innovations, $\{\nu_1, \ldots, \nu_L\}$,

and then, the estimator, $\widehat{x}_{k/L}$, based on the observations can be expressed as a linear combination of the innovations,

$$\widehat{x}_{k/L} = \sum_{i=1}^{L} S_{k,i}\Pi_i^{-1}\nu_i$$

where $S_{k,i} = E[x_k\nu_i^T]$ and Π_i is the innovation covariance matrix.

3 Filtering and fixed-point smoothing algorithms

Theorem 1. Under the hypotheses (I)-(IV) set out in Section 2, the filter of the signal is calculated as

$$\widehat{x}_{k/k} = \alpha_k O_k, \quad k \geq 1$$

where the vectors O_k are recursively calculated as

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

with

$$\begin{aligned} J_k &= p_k[\beta_k^T - r_{k-1}\alpha_k^T], \\ r_k &= r_{k-1} + J_k\Pi_k^{-1}J_k^T, \quad k \geq 1; \quad r_0 = 0. \end{aligned}$$

The innovation, ν_k, and its covariance matrix, Π_k, are given by

$$\begin{aligned} \nu_k &= y_k - p_k\alpha_k O_{k-1} - (1-p_k)y_{k-1}, \quad k > 1; \quad \nu_1 = y_1 \\ \Pi_k &= q_k - p_k^2\alpha_k r_{k-1}\alpha_k^T - p_k(1-p_k)[\alpha_k l_{k-1} + l_{k-1}^T\alpha_k^T] - (1-p_k)^2 q_{k-1} \end{aligned}$$

where

$$\begin{aligned} q_k &= p_k[\alpha_k\beta_k^T + R_k] + (1-p_k)q_{k-1}, \quad k \geq 1; \quad q_0 = 0 \\ l_k &= J_k + p_k r_{k-1}\alpha_k^T + (1-p_k)l_{k-1}, \quad k \geq 1; \quad l_0 = 0. \end{aligned}$$

Theorem 2. Under the hypotheses (I)-(IV) set out in Section 2, the fixed-point smoother of the signal is given by

$$\widehat{x}_{k/L} = \widehat{x}_{k/L-1} + S_{k,L}\Pi_L^{-1}\nu_L, \quad L > k$$

with initial condition $\widehat{x}_{k/k}$, given in Theorem 1.

The smoothing gain, $S_{k,L}$, is

$$S_{k,L} = p_L[\beta_k - F_{k,L-1}]\alpha_L^T, \quad L > k$$

where

$$F_{k,L} = F_{k,L-1} + S_{k,L}\Pi_L^{-1}J_L^T, \quad L > k; \quad F_{k,k} = \alpha_k r_k.$$

Finally, expressions for the filtering and fixed-point smoothing error variances are given, respectively, by

$$\begin{aligned} P_{k/k} &= \alpha_k[\beta_k^T - r_k\alpha_k^T], \quad k \geq 1, \\ P_{k/L} &= P_{k/L-1} - S_{k,L}\Pi_L^{-1}S_{k,L}^T, \quad L > k. \end{aligned}$$

4 Simulation example

The proposed algorithms are applied to a simulation example to illustrate its effectiveness. Consider a zero-mean scalar signal x_k with known covariance function which is expressed according to hypothesis (I), being $\alpha_k = 1.025641 \times 0.95^k$ and $\beta_s = 0.95^{-s}$. The observations of $z_k = x_k + v_k$ are modeled as (1), where $\{\xi_k,\ k \geq 1\}$ are independent Bernoulli random variables with $P[\xi_k = 1] = p, \forall k \geq 1$ and $\{v_k,\ k \geq 1\}$ is white noise with zero mean and $R_k = 0.9$. The filtering and smoothing error variances for different probabilities, $p = 0.2, 0.5$ and 0.8, are displayed in Figure 1. This figure shows, on the one hand, that as the probability increases, the error variances decrease and so, the performance of estimators is better. On the other hand, it also is observed that for each p the smoothing error variances, $P_{k/k+2}$, are less than the corresponding filtering ones, $P_{k/k}$, and, as was expected, the better estimations are obtained for the fixed-point smoother.

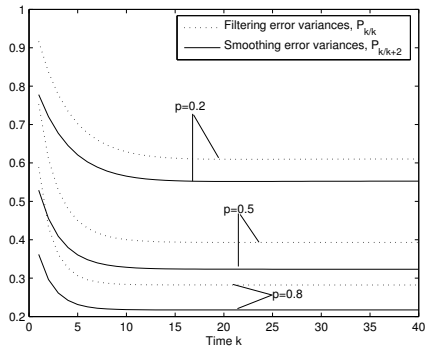


FIGURE 1. Filtering and fixed-point smoothing error variances

Acknowledgments: This work was partially supported by "Ministerio de Educación y Ciencia" under contract MTM2008-05567.

References

Nakamori, S., García-Ligero, M.J., Hermoso-Carazo, A. and Linares-Pérez, J. (2006). Derivation of fixed-interval smoothing algorithm using covariance information in distributed parameter systems. *Applied Mathematics and Computation*, **176**, 662-672.

Sahebsara, M., Chen, T. and Shah, L. (2007). Optimal H_2 filtering with random sensor delay, multiple packet dropout and uncertain observations. *International Journal of Control*, **80**, 292-301.

Sun, S., Xie, L., Xiao, W., Soh, Y. Ch. (2008). Optimal linear estimation for systems with multiple packet dropouts. *Automatica*, **44**, 1333-1342.

Multivariate Modelling of a Monotone Disease Process in the Presence of Misclassification

María José García-Zattera[1,2,†], Alejandro Jara[1], Emmanuel Lesaffre[2,3], Guillermo Marshall[1]

[1] Department of Statistics, Pontificia Universidad Católica de Chile, Casilla 306, 22 Santiago, Chile
[2] L-BioStat, Katholieke Universiteit Leuven, Kapucijnenvoer 35, Block D, Bus 7001, B-3000 Leuven, Belgium
[3] Department of Biostatistics, Erasmus Medical Center, Dr Molewaterplein 50, 3015 GE Rotterdam, the Netherlands
† Email: mjgarcia@uc.cl

Abstract: Motivated by a longitudinal oral health study, the Signal-Tandmobiel®, we propose a multivariate binary inhomogeneous Markov model in which correlated true response variables are subject to an unconstrained misclassification process and follow a monotone behavior. Conditionally specified logistic regression models are used to relate the multivariate baseline distribution and the elements of the transition matrices of the unobserved process to the available covariates. The misclassification model considers the existence of different classifiers for each subject across time.

Keywords: Multivariate Binary Data; Monotone Processes; Misclassification; Incidences Estimation; Hidden Markov Model.

1 Introduction

The motivation for this work comes from data gathered in a longitudinal oral health study conducted in Flanders (Belgium) between 1996 and 2001, the Signal-Tandmobiel® (ST) study (Vanobbergen *et al.*, 2000). For this project, 4,468 children were examined on a yearly basis during their primary school time (between 7 and 12 years of age) by one of sixteen dental examiners. The purpose of the investigation is to examine the association of several potential risk factors with the prevalence and incidence of caries experience (CE), which is defined as a binary variable indicating whether a tooth is decayed at d_3 level, missing or filled due to caries. The assessment of risk factors under this setting involves the analysis of a misclassified multivariate monotone binary process since: i) CE, as previously defined, is a progressive or monotone disease because teeth cannot alternate between the presence or absence of CE once CE is observed over time, ii) events on

teeth of the same child are dependent, and iii) the examiners scoring may not perfectly reflect the tooth's true condition and the presence of CE can be miss-diagnosed.

In this work we extent recent developments on univariate hidden binary monotone Markov models (García-Zattera *et al.*, 2010) to provide a framework for the analysis of multivariate hidden monotone process as a function of covariates. The rest of the manuscript is organized as follows. In Section 2 we introduce the proposed model. Section 3 presents the analysis of our motivating problem. A final discussion section concludes the manuscript.

2 The model

Suppose that J teeth are examined on subject i, $i = 1, \ldots, I$, at time points $t_{(i,k)}$, $k = 1, \ldots, K$. Let $Y_{(i,j,k)}$ be the true unobserved binary response for tooth j of subject i at time $t_{(i,k)}$ and denote the J-dimensional vector of true responses for all teeth of subject i at time $t_{(i,k)}$ by $\boldsymbol{Y_{(i,k)}} = (Y_{(i,1,k)}, \ldots, Y_{(i,J,k)})'$. Let $\boldsymbol{x_{(i,j,k)}}$ be a p-dimensional vector of covariates for subject i and tooth j at examination k. We assume that the vectors $\boldsymbol{Y_{(i,k)}}$ follow a monotone inhomogeneous first-order Markov process. In order to relate the covariates to the initial distribution and to the elements of the transition matrices in a population-average manner and taking into account the association among the responses of the same subject, conditionally specified logistic regression models (Joe and Liu, 1996; García-Zattera et al., 2007) are used. Specifically, we assume that

$$\Pr(\boldsymbol{Y_{(i,1)}} = \boldsymbol{y} \mid \boldsymbol{X_{(i,1)}}) \propto \exp\left\{ \sum_{j=1}^{J} \left(\boldsymbol{x_{(i,j,1)}}' \boldsymbol{\beta_j^P} \right) y_j + \sum_{1 \le j < l \le J} \gamma_{jl}^P y_j y_l \right\},$$

and, for $k = 2, \ldots, K$,

$$\begin{aligned} \Pr(\boldsymbol{Y_{(i,k)}} = \boldsymbol{y^1} \mid \boldsymbol{Y_{(i,k-1)}} = \boldsymbol{y^0}, \boldsymbol{Z_{(i,k)}}) \quad \propto \quad & \exp\left\{ \sum_{j=1}^{J} \left(\boldsymbol{z_{(i,j,k)}}' \boldsymbol{\beta_j^I} \right) y_j^1 (1 - y_j^0) + \right. \\ & \sum_{1 \le j < l \le J} \gamma_{jl}^I y_j^1 y_l^1 (1 - y_j^0 y_l^0) + \\ & \left. \sum_{j=1}^{J} \sum_{j \ne l}^{J} \alpha_{jl}^I \; y_j^0 y_l^1 (1 - y_j^0 y_l^0) \right\}, \end{aligned} \tag{1}$$

where $\boldsymbol{y} \in \{0,1\}^J$, $\boldsymbol{y^1} \in \mathcal{B}(\boldsymbol{y^0}) \subset \{0,1\}^J$, with $\mathcal{B}(\boldsymbol{y^0})$ being an admissible set, $\boldsymbol{z_{(i,j,k)}} = (\boldsymbol{x_{(i,j,k-1)}}', t_{(i,k)} - t_{(i,k-1)})$, $\boldsymbol{\beta^P}$ and $\boldsymbol{\beta^I}$ are vectors of conditionally specified logistic regression models associated to the initial distribution and transition matrices, respectively, $\boldsymbol{\gamma^P}$ and $\boldsymbol{\gamma^I}$ are vectors of within-time conditional log-odds ratio parameters for the initial distribution and transition matrices, respectively, and $\boldsymbol{\alpha^I}$ is a vector of across-time

conditional log-odds ratios. We show that expression (1) defines a proper probability model for each row in the Markovian transition matrices.

We assume that the response variables $Y_{(i,j,k)}$ are prone to misclassification. Let $Y^*_{(i,j,k)}$ be the corrupted observed binary response for tooth j of subject i at time $t_{(i,k)}$. Here we suppose that the scoring is performed by Q examiners. Denote by $\xi_{i,k} \in \{1, \ldots, Q\}$ the indicator variable of examiner that scores all teeth of subject i at time $t_{(i,k)}$.
In an initial version of the model (M_1) we assume that the scoring behavior of the examiners is the same across the study and teeth. Let $\tau_{e,11}$ and $\tau_{e,00}$ be the sensitivity and specificity, respectively, for examiner e, $e = 1, \ldots, Q$. The misclassification model assumes that $\Pr\left(Y^*_{(i,j,k)} = 1 \mid Y_{(i,j,k)} = 1\right) = \tau_{\xi_{i,k},11}$ and $\Pr\left(Y^*_{(i,j,k)} = 0 \mid Y_{(i,j,k)} = 0\right) = \tau_{\xi_{i,k},00}$. In a second version of the model (M_2), we assume that the scoring behavior of the examiner is tooth-specific, but the same across the study; i.e., here $\tau^j_{e,11}$ and $\tau^j_{e,00}$, $j = 1, \ldots, J$, denote the tooth-specific sensitivity and specificity, respectively, for examiner e. By using simulated data, we show that, under the setting of the motivating problem, the restriction $\tau_{e,00} + \tau_{e,11} > 1$ for M_1 and $\tau^j_{e,00} + \tau^j_{e,11} > 1$ for M_2, along with restrictions on the rank of the design matrices, are sufficient conditions in order to ensure parameters identifiability.

3 The analysis of the Signal-Tandmobiel® data

We implement Bayesian versions of the models and assume independent $N(0, 10^3)$ priors for the coordinates in $\boldsymbol{\beta^P}$, $\boldsymbol{\beta^I}$, $\boldsymbol{\gamma^P}$, $\boldsymbol{\gamma^I}$ and $\boldsymbol{\alpha^I}$. For the misclassification parameters, independent uniform distributions are assumed, under the restriction $\tau_{e,00} + \tau_{e,11} > 1$ for M_1 and $\tau^j_{e,00} + \tau^j_{e,11} > 1$ for M_2. We fit the models to the four permanent first molars, teeth 16, 26 on the maxilla (upper quadrants), and teeth 36 and 46 on the mandible (lower quadrants). We evaluated the effect of the age at start of brushing (in years), the number of between-meal snacks (two or less than two a day vs more than two a day), the geographical location, and the age (in years) on the prevalence and incidencees of CE. Model comparison was performed using the pseudo Bayes factor (PsBF) (Geisser and Eddy, 1979).

The $2 \times \log_{10}$PsBF for M_1 versus M_2 was 40.6, suggesting no evidence for the hypothesis of tooth-specific misclassification parameters for each examiner. Therefore, we only report the results arising from M_1. Table 1 shows the posterior mean and 95% highest posterior density (95% HPD) credible intervals for the conditionally specified logistic regression coefficients. The results suggest that the older the child, the higher the prevalence of CE, and that the later the child starts brushing or the higher the number of between-meal snacks, the higher the incidence of CE.

TABLE 1. Posterior mean and 95% highest posterior density (95% HPD) credible intervals, for the conditionally specified logistic regression coefficients associated to the prevalence and incidences for CE in permanent first molars.

	Prevalence		**Incidences**	
	Posterior Mean	**95%HPD**	**Posterior Mean**	**95%HPD**
Intercept T16	-6.16	(-8.78 ; -3.54)	-4.62	(-5.53 ; -3.59)
Intercept T26	-5.62	(-8.20 ; -3.13)	-4.33	(-5.25 ; -3.37)
Intercept T36	-5.75	(-8.36 ; -3.26)	-4.65	(-5.62 ; -3.63)
Intercept T46	-5.58	(-8.07 ; -3.02)	-4.21	(-5.12 ; -3.24)
Age Start Brushing	0.09	(-0.02 ; 0.20)	0.09	(0.04 ; 0.13)
Age	0.35	(0.00 ; 0.69)	-0.03	(-0.08 ; 0.01)
Meals	0.15	(-0.11 ; 0.41)	0.15	(0.05 ; 0.27)
x-ordinate	0.00	(-0.24 ; 0.25)	0.07	(-0.04 ; 0.17)
y-ordinate	-0.60	(-1.43 ; 0.18)	-0.16	(-0.50 ; 0.17)
Years Between Exam.	-	-	0.39	(-0.08 ; 0.82)

Posterior mean and 95% HPD credible intervals for the association parameters are shown in Table 2. The posterior inferences for the within-time conditional log-odds suggest a high positive association in the presence of CE between symmetrically opponent molars and right vertically opponent molars (maxilla versus mandible) at the age of 7. At this age, a non-significant conditional association was found between diagonally opponent teeth. High positive within-time conditional associations were found between symmetrically, right vertically opponent molars and diagonally opponent teeth as the process evolves.

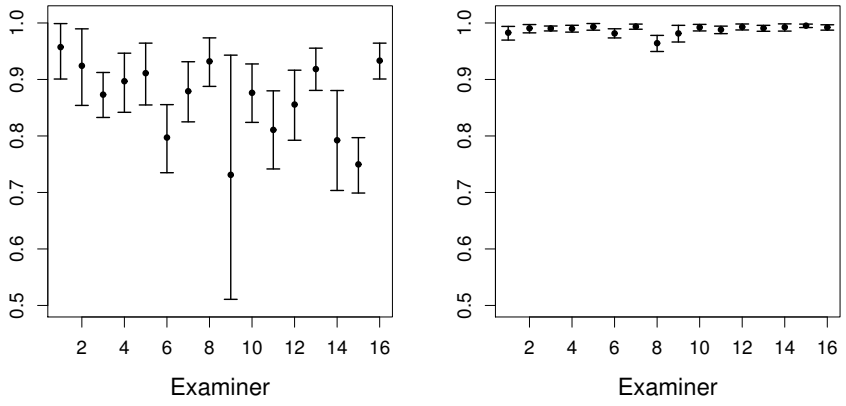
The posterior inference for the across-time odds ratio parameters suggest significant and negative associations between symmetrically and diagonally opponent molars. These results suggest that the probability of developing caries on a tooth is higher when a symmetrically or diagonally opponent molar is affected at the same time but sound at the previous examination, than when it was affected in the previous examination. This result can be explained by the fact that once a tooth is affected by caries, it is probably treated and the infection is no longer spreading in the next examination.

Finally, Figure 1 shows the posterior mean and 95% HPD credible intervals for the sensitivity and specificity for each examiner. The results suggest a greater variability in the sensitivity than in the specificity estimates, which can be explained by the low prevalence and incidences of CE. All examiners showed a sensitive greater than 0.7, with the exception of one examiner who showed a rather poor scoring behavior. The latter result is explained by the lower information available for this examiner. In fact, this examiner was only involved in the first two years of the ST study. The posterior mean for the specificity parameters were higher than 0.96 for all examiners.

TABLE 2. Posterior mean and 95% highest posterior density (95% HPD) credible intervals of conditional odds ratios for CE in permanent first molars.

		Prevalence		Incidences	
		Posterior Mean	**95%HPD**	**Posterior Mean**	**95%HPD**
Within Time Association Parameters	$\gamma_{16,26}$	3.95	(2.84 ; 5.13)	3.86	(3.18 ; 4.55)
	$\gamma_{16,36}$	0.59	(-1.74 ; 2.92)	2.35	(1.25 ; 3.38)
	$\gamma_{16,46}$	2.30	(0.27 ; 4.07)	1.11	(0.19 ; 2.09)
	$\gamma_{26,36}$	1.57	(-1.05 ; 3.72)	0.63	(-1.94 ; 0.78)
	$\gamma_{26,46}$	-0.21	(-2.52 ; 2.34)	2.11	(1.15 ; 2.99)
	$\gamma_{36,46}$	2.63	(1.36 ; 3.91)	3.71	(3.03 ; 4.39)
Previous Time Association Parameters	$\alpha_{16,26}$	-	-	-2.82	(-3.84 ; -1.80)
	$\alpha_{16,36}$	-	-	-1.46	(-2.75 ; -0.22)
	$\alpha_{16,46}$	-	-	-0.44	(-1.53 ; 0.74)
	$\alpha_{26,16}$	-	-	-1.91	(-2.91 ; -0.95)
	$\alpha_{26,36}$	-	-	0.78	(-0.76 ; 2.28)
	$\alpha_{26,46}$	-	-	-1.53	(-2.59 ; -0.39)
	$\alpha_{36,16}$	-	-	-2.08	(-3.33 ; -0.82)
	$\alpha_{36,26}$	-	-	1.06	(-0.54 ; 2.55)
	$\alpha_{36,46}$	-	-	-2.48	(-3.53 ; -1.49)
	$\alpha_{46,16}$	-	-	-0.48	(-1.55 ; 0.59)
	$\alpha_{46,26}$	-	-	-1.01	(-2.14 ; 0.06)
	$\alpha_{46,36}$	-	-	-1.22	(-1.97 ; -0.47)

FIGURE 1. Posterior mean and 95% highest posterior density credible intervals for examiner's sensitivity (panel a) and specificity (panel b).



(a) (b)

4 Concluding Remarks

We have proposed a multivariate hidden Markov model for monotone binary processes, which describes the relationships between covariates and the prevalence and incidences in a population-average fashion, and where different classifiers are present. The association between the responses, is taken into account by within- and across-time conditional odds ratios. An advantage of the proposed model is that the parameters can be estimated without external information on the misclassification parameters. Results from simulation studies suggest that even under the use of uniform priors on the misclassification parameters, unbiased and precise estimates of the parameters can be obtained.

References

García-Zattera, M.J., Jara, A., Lesaffre, E. and Declerck D. (2007). Conditional independence of multivariate binary data with an application in caries research. *Computational Statistics and Data Analysis*, **51: 3223–3234**.

García-Zattera, M.J., Mutsvari, T., Jara, A., Declerck D. and Lesaffre, E. (2010). Correcting for misclassification for a monotone disease process with an application in dental research. *Statistics in Medicine*, **Accepted for publication**.

Geisser, S. and Eddy, W. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, **74: 153–160**.

Joe, H. and Liu, Y. (1996). A model for a multivariate binary response with covariates based on compatible conditionally specified logistic regressions. *Statistics and Probability Letters*, **31: 113–120**.

Vanobbergen J., Martens L., Lesaffre E. and Declerck D. (2000). The Signal-Tandmobiel project, a longitudinal intervention health promotion study in Flanders (Belgium): baseline and first year results. *European Journal of Paediatric Dentistry*, **2: 87–96**.

Dynamic programming versus graph cut algorithms for fitting non-parametric models to image data

C. A. Glasbey[1]

[1] Biomathematics & Statistics Scotland, King's Buildings, Edinburgh EH9 3JZ, Scotland

Abstract: Dynamic programming and graph cut algorithms can, in some cases, find globally optimal fits of non-parametric models to image data, for restoration, segmentation and template matching. In this paper we compare conditions and results for the two methods, illustrated by restoration of a synthetic aperture radar (SAR) image.

Keywords: Image Restoration; Maximum a Posteriori Estimator; Markov Random Field; Penalised Likelihood; Synthetic Aperture Radar.

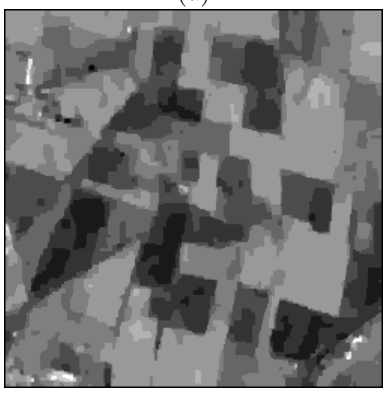
1 Introduction

Non-parametric models are fit to image data for many reasons, including restoration, segmentation and template matching. For example, synthetic aperture radar (SAR) is a form of remote sensing in which data have a large noise component and therefore need smoothing, or restoring, to simplify interpretation. To illustrate, Figure 1(a) shows a log-transformed SAR image of an area near Thetford in England, obtained by plane in the Maestro-1 campaign. A pattern of agricultural fields can be discerned but, although we would expect the true signal to be approximately constant within each field, there is considerable speckle. SAR image restoration can be formulated as non-parametric smoothing:

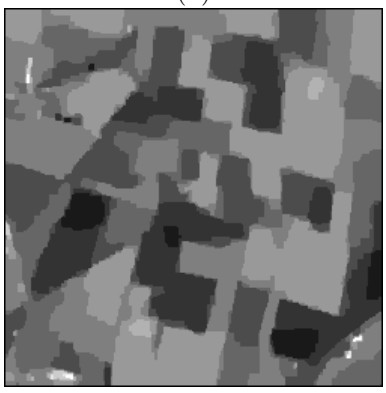
$$\hat{\beta} = \arg\min_{\beta \in \mathcal{B}} \left\{ \sum_i (y_i - \beta_i)^2 + \sum_{\|i-j\|=1} \lambda |\beta_i - \beta_j| \right\}. \tag{1}$$

Here, y is the log-transformed SAR data, a $P = 250 \times 250$ array of scalars (rescaled to lie in the range 0 and 255) indexed by i, a 2-dimensional integer vector, and β is the scalar 2-dimensional array of restored values, an estimate of the true SAR signal, constrained to a finite set $\mathcal{B}$ which, to speed-up computations, we restrict to a small set of values $\{65, 75, \ldots, 165\}$. Summations are over all possible values of i and of i and j, and $\|.\|$ denotes the Euclidean metric. So, we seek the least-squares restoration of y, subject to a penalty for lack of smoothness in β, specified by an absolute difference penalty with a first-order neighbourhood, and scaled by λ, a non-negative

(a)



(b)



(c)

FIGURE 1. Synthetic aperture radar (SAR) image: (a) log-transformed data downloadable from `ftp://ftp.bioss.sari.ac.uk/pub/chris/warping/`; (b) restoration using iterated dynamic programming (IDP); (c) restoration using graph cut algorithm.

constant whose magnitude determines the smoothness of the fit. By using this form of penalty, we impose smoothness on the restoration while tolerating step changes that we anticipate at agricultural field boundaries, and we set $\lambda = 50$, determined by eye to produce realistic restorations. Although the objective function in (1) is convex, this is a pathological optimisation problem for which gradient descent algorithms are unsuccessful (Kunsch, 1994).

More generally, image restoration, segmentation and template matching can often be formulated, using either a Bayesian or penalised likelihood framework, as

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{B}} \left\{ \sum_i f(y, \beta_i) + \sum_{\|i-j\|=1} g(\|\beta_i - \beta_j\|) \right\}. \tag{2}$$

Here, β is the I-dimensional array of B-dimensional vectors that specifies the model fit, indexed by i, an I-dimensional integer vector, f is the measure of model fit to the data, derived from the negative log-likelihood of y, and g is either an empirical term penalising lack of smoothness in β, or the log-prior of β, a Markov Random Field (MRF) with first-order neighbourhood. We restrict β to a finite set $\mathcal{B} = \{1, 2, \ldots, n\}^B$, though this is not overly restrictive since any desired level of resolution can be achieved by setting n sufficiently large and rescaling y. There are many choices for f and g, depending on the application. For image restoration, one option is $f = (y_i - \beta_i)^2$ as in (1), whereas for warping, or matching, to template μ, $f = (y_i - \mu_{i+\beta_i})^2$, and for segmentation we could use $f = y^*_{\beta_i}$, where y^* denotes transformed data (Glasbey and Young, 2002). The penalty term may be convex, such as $g(z) = \lambda z^2$ or $\lambda|z|$, as used in (1), or non-convex, such as the indicator function.

Solving (1) or (2) is computationally challenging for two reasons: image data sets are large and such objective functions are prone to local optima. However, if certain conditions apply, dynamic programming and graph cut algorithms can be used, and they are among that rare class of optimisation algorithms, those that are both fast and global!

2 Algorithms

Dynamic programming (DP) is an elegant method for finding the global solution to (2), but only if $I = 1$ (i.e. β is a one-dimensional array), using a sequential algorithm. Cases include 1-D image warping, also termed dynamic time warping, and finding 1-D boundaries to segment 2-D images, by finding a path between opposite sides of an image, with β_i specifying the row location of the boundary in column i (Glasbey and Young, 2002).

DP cannot be used to restore the SAR image by solving (1) because $I = 2$. However, Glasbey (2009) proposed generalisations of DP to solve higher

dimensional problems, but without the guarantee of global optimality; the simplest being iterated dynamic programming (IDP). IDP can be initialised by applying DP separately to each image column, and subsequently DP is applied alternately to all rows and columns, taking into account neighbouring values of β. After 26 iterations, which took 1.2sec of CPU time on a single core of a 3.2Ghz AMD Opteron processor, β converged to an approximation to the maximum *a posteriori* (MAP) or maximum penalised likelihood estimator shown in Figure 1(b), with a minimised value of the objective function of $388P$.

Graph cut (GC) algorithms can also be used to find the global solution to (2), provided $B = 1$ (i.e. β is a scalar array) and g is a convex function. GC reformulates the optimisation problem as finding the maximum flow through a network from a source to a sink, or equivalently, the minimum-cost subset of edges which disconnect the source from the sink. We create an $I + 1$-dimensional array of $P \times (n + 1)$ nodes, indexed by (i, β), with edges and directional flow capacities specified by:

$$\begin{aligned} C\{\text{source} \to (i,1)\} &= \infty \\ C\{(i,n+1) \to \text{sink}\} &= \infty \\ C\{(i,\beta) \to (i,\beta+1)\} &= f(y,\beta) \\ C\{(i,\beta) \to (j,\beta) \mid \|i-j\| = 1\} &= g(1) \\ C\{(i,\beta) \to (j,\beta-k) \mid \|i-j\| = 1\} &= \{g(k+1) - 2g(k) + g(k-1)\} \text{ for } k \geq 1, \end{aligned}$$

where, without loss of generality we assume that $g(0) = 0$. The original, Ford-Fulkerson algorithm solves the flow problem by repeatedly finding a path from source to sink with spare capacity and maximising flow along this path. Grieg et al (1989) were the first to use GC in image analysis, to restore binary images. More recently, Boykov and co-workers (see, for example, Boykov and Kolmogorov, 2004) have extended GC to a broader class of image problems, and shown that a different type of GC algorithm, termed 'push-relabel', is more effective for these applications.

SAR image restoration (1) satisfies the conditions for GC. Further, because $g(z) = \lambda|z|$, the last, large set of edges is not needed, though there are still 750,000 nodes and 4.6 million edges. Boykov's implementation of the GC algorithm (available at `http://vision.csd.uwo.ca/code/`) took 1.1sec of CPU time to find the globally optimal result. However, there are many alternative GC algorithms, and speed can vary enormously. For example, another public domain implementation, NETFLO (Nijenhuis and Wilf, 1978), using a Ford-Fulkerson type algorithm, took 10^3 more time (21 minutes) to reach the same solution. The MAP or maximum penalised likelihood estimator is shown in Figure 1(c), with a minimised value of the objective function of $374P$. We see that the agricultural fields are more clearly identified than in Figure 1(b), presumably because we have the global optimum rather than simply a local one.

3 Discussion

DP and GC can each fit non-parametric models to image data by finding the global solution to (2) provided certain conditions apply. For DP, the condition is $I = 1$, whereas for GC, $B = 1$ and g must be convex. As we have seen, GC can restore 2-D images, or any higher dimension, for convex choices of g, whereas DP can only restore 1-D images, and for higher dimensions we have to resort to approximations such as IDP, though they have the advantage of permitting non-convex g. Note that, for the special case in (1), Kovac and Smith (2009) have shown that a taut-string type algorithm can also find the global optimum. In unpublished work, we have also compared IDP and GC for finding 2-D surfaces to segment 3-D images (i.e. $B = 1$, $I = 2$), which DP cannot achieve. However, if instead the model required us to find a 1-D path across a 3-D image ($B = 2$, $I = 1$) then DP could do it but GC would not be able to. DP and GC can both be used in 1-D warping ($B = 1$, $I = 1$), but only GC can perform so-called $1\frac{1}{2}$-D warping, which occurs, for example, when matching stereo image pairs, where no warping is needed between rows (i.e. $B = 1$, $I = 2$). However, neither DP nor GC can perform 2-D or 3-D warping.
Overall, GC is a much more complicated algorithm than DP, and its many variants can have an enormous range of performance speeds. DP is easier to generalise, for example to MRFs with higher-order neighbourhoods and to find local optima using IDP even when DP cannot be used. GC can also be adapted, to find local optima when conditions for global optimality are not satisfied, but then other competitors exist, such as 'loopy belief propagation' and 'tree-reweighted message passing' (Szeliski et al, 2008).

Acknowledgments: I am grateful to my colleague Alec Mann for assistance in using Boykov's C++ algorithm. The work was funded by the Scottish Government.

References

Boykov, Y., and Kolmogorov, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26**, 1124-1137.

Glasbey, C.A. (2009). Two-dimensional generalisations of dynamic programming for image analysis. *Statistics and Computing*, **19**, 49-56.

Glasbey, C.A., and Young, M.J. (2002). Maximum *a posteriori* estimation of image boundaries by dynamic programming. *Applied Statistics*, **51**, 209-221.

Greig, D.M., Porteous, B.T., and Seheult, A.H. (1989). Exact maximum *a posteriori* estimation for binary images. *Journal of the Royal Statistical Society, Series B*, **51**, 271-279.

Kovac, A., and Smith, A.D.A.C. (2009). Regression on a graph. (Available on arXiv, at http://arxiv.org/abs/0911.1928v1)

Kunsch, H.R. (1994). Robust priors for smoothing and image restoration. *Annals of the Institute of Statistical Mathematics*, **46**, 1-19.

Nijenhuis, A., and Wilf, H.S. (1978). *Combinatorial Algorithms* (2nd edition). London: Academic Press

Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., and Rother, C. (2008). A comparative study of energy minimization methods for Markov random fields with smoothness-based priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30**, 1068-1080.

Estimating Thermocline Depth In Lakes

FM Hamzah[1], EM Scott[1], CA Ferguson[1], S Waldron[2], M O'Hare[3], CE Adams [3]

[1] Department of Statistics, University of Glasgow, G12 8QQ, UK
[2] Department of Geographical and Earth Sciences University of Glasgow, G12 8QQ, UK
[3] The Scottish Centre for Ecology and the Natural Environment, University of Glasgow, Rowardennan, Drymen, Glasgow, G63 0AW, UK

Abstract: The thermocline in a lake is an imaginary plane located at the depth where the rate of temperature decrease (temperature gradient) in the temperature profile is at a maximum; this phenomenon can greatly affect the biological and chemical processes in lakes. The temperature profile over 11 different depths in the north basin of Loch Lomond, Scotland, is explored here to investigate thermocline development in the lake throughout the year. A changepoint regression method is utilised as one approach to estimate the thermocline position over time. For the Loch Lomond data, there is evidence of the thermocline at shallow depths in warmer months.

Keywords: Changepoint regression; Thermocline development; Loch Lomond

1 Introduction

A well-known natural phenomenon that occurs within most lakes during the summer is the development of temperature stratification. According to widely accepted limnological convention, the thermocline is an imaginary plane located at the depth where the rate of temperature decrease in the temperature profile is maximum. The thermocline divides warmer waters in the top layer from colder waters lying below and as a consequence, the lake is divided into two strata with different biological and chemical features. For further information please see Wetzel, 2001.
This natural phenomenon was investigated to determine the thermocline development in Loch Lomond, Scotland, over time. Here, 3-hourly water temperature (^{o}C) measurements from 1 September 2002 to 31 August 2003, which were recorded at 11 different depths at Cailness in the north basin of Loch Lomond, are investigated for evidence of the formation of the thermocline. The data were collected as part of the Eurolakes project.

2 Methods

A contour plot of temperature across depth and year is used to subjectively investigate the temperature profile in the water column. In order to estimate the thermocline depth, a method is required that estimates positions of rapid change in temperature automatically for the 2920 temperature profiles. Initially, a changepoint regression method is used to estimate the thermocline depth.

2.1 Changepoint regression method

Changepoint regression, see for example, Krisnaih and Miao (1988) and Julious (2001) can be applied to a data series where the regression slope is not constant but could change rapidly at a given point. Changepoints are of primary interest here as they may indicate rapid change in temperature with depth, and hence such points might identify the thermocline formation and depth.
The estimation of the parameters in the model is straightforward if the location of the thermocline is known. However, if it is not known, the changepoint must be estimated. In addition, the problem is no longer linear and numerical optimization is used to estimate the parameters (Julious, 2001). At each time point, for any interval (X_0, X_1) on the real line where X_0 and X_1 are depths in this context, the problem is defined as:

$$f(x_i) = \begin{cases} f_1(x_i;\beta_1); X_0 \leq x_i \leq \tau, \\ f_2(x_i;\beta_2); \tau \leq x_i \leq X_1 \end{cases}$$

where $f(x_i)$ is the temperature at depth x_i and τ is the changepoint such that $f_1(x_i;\beta_1) = f_1(x_i;\beta_2)$ i.e. the slope of the relationship between $f(x)$ and x is β_1 until τ on the x-axis and the slope is β_2 thereafter. There is no discontinuity between the two regression lines.

3 Results

3.1 Exploratory Analysis

The contour plot of temperature in Figure 1 with a contour step of 1^0C, clearly shows that the temperature in the water column is homogeneous between December 2002 and March 2003, but greater changes are evident in the other months. Areas of sharp temperature gradients, which are shown by several contours close to each other, may indicate the position of the thermocline. This feature can be clearly seen from September to November 2002 and May to August 2003. However, the temperature profiles with depth appear to be different in each of these two time periods.
Changepoint regression (Julious, 2001) has been used to formally investigate the thermocline depth.

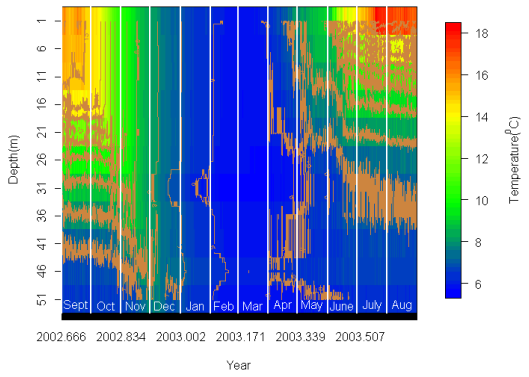


FIGURE 1. Contour plot of temperature across depth and year

3.2 Changepoint regression method

Figure 2 exhibits a time series plot of the estimated depth at which the maximum change in slope occurs from 1 September 2002 to 31 August 2003 using a changepoint regression method. The plot highlights changepoints from October to December 2002 and April to August 2003. There were no changepoints identified for September and early October, and between January and March 2003. The results in September and October 2002 contradict the characteristics of the temperature profiles in the exploratory analysis, which suggest evidence of the thermocline formation in September and October 2002. This is a result of the fact that the changepoint regression method used here cannot detect the multiple changepoints evident in these months. The changepoints identified in November and December 2002 are more likely to be other features in the temperature profile since the thermocline can only occur in the warmer months. In 2003 there is evidence of the thermocline appearing in April in the middle of the water column and moving upwards through the water column (although there is no clear temporal pattern), between June and August.

4 Discussion and Future Work

Changepoint regression has been used to estimate the thermocline depth in the north basin of Loch Lomond. While results appear realistic for summer 2003, the method was unable to estimate the thermocline depths when there was more than one changepoint in September and October 2002. Mathematically, the thermocline depth is the inflection point of the temperature curve; the depth where the temperature gradient changes concavity. This

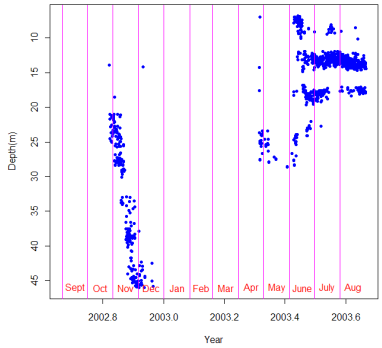


FIGURE 2. The estimated change points from 1 September 2002 to 31 August 2003

definition appears to be appropriate for the temperature profiles in September and October 2002, Figure 1, which have two changepoints as a result of the curves changing concavity. The inflection point can be estimated by fitting a nonparametric regression to each profile and calculating the second derivative at each point of interest along the curve. Preliminary results suggest that the inflection points (as an indicator of the thermocline depth) lie between 17 and 39m for September and October 2002.

The current results will be compared with the data from nearby location in Loch Lomond, which consists of hourly temperature measurements at 11 different depths, from 17 April 2008 to 27 May 2009. This will enable us to investigate if the pattern observed here is typical of a yearly process in Loch Lomond.

Acknowledgments: Firdaus Mohamad Hamzah, funded by the Malaysian Government and National University of Malaysia (UKM).

References

Julious, S.A. (2001). Inference and estimation in a changepoint regression problem. *Journal of the Royal Statistical Society, Series D*, **50**.

Krisnaiah, P.K. and Miao, B.Q. (1988). *Review about estimation of changepoints. In Handbook of Statistics.* Amsterdam: North-Holland.

Wetzel, R.G. (2001). *Limnology: lake and river ecosystems.* San Diego; London; Academic Press

Multilevel latent-class modelling of patient casemix

Wendy J. Harrison[1], Robert M. West[1], Amy Downing[1], David Forman[12], Mark S. Gilthorpe[1]

[1] Centre for Epidemiology & Biostatistics, School of Medicine, University of Leeds, Leeds, LS2 9JT, UK
[2] Northern & Yorkshire Cancer Registry & Information Service, Bexley Wing, St James's Institute of Oncology, Leeds, LS9 7TF, UK

Abstract: Casemix adjustment and institutional comparison are demonstrated through multilevel latent-class regression with an extensive colorectal cancer dataset. Distributional assumptions at the upper level are more appropriate and uncertainty of patient assignment to casemix is accomodated.

Keywords: Multilevel latent-class regression; Casemix adjustment.

1 Background

Survival from cancer may vary by place of diagnosis and treatment centre (Trust), stage at diagnosis, age at diagnosis, sex, and socioeconomic background (SEB). Some Trusts perform better than others in terms of their average survival rates and this might be due to casemix, though outcome differences might also be due to genuine underlying differences in the effectiveness of healthcare organisations. Much interest lies in identifying good and poor performing healthcare providers. It is important to account for patient casemix when evaluating institutional performance and there are several strategies. Regression models are a traditional and well-documented approach, where variables relating to patient characteristics are modelled to adjust the outcome in relation to these. Matching, stratification, or propensity-score analysis, may also be used, though such techniques can make untestable assumptions and never account for unmeasured variables. No casemix adjustment strategy will eliminate bias due to unmeasured differences entirely, thus accommodating patient variation via the measured variables only is over-simplistic: models need to accommodate uncertainty associated with patient casemix characteristics. Using routine data, this study explores the utility of multilevel latent-class (MLLC) modelling to adjust for patient casemix characteristics, their associated uncertainty, and Trust variation, in order to rank Trust performance.
This paper develops the muiltilevel latent-class analysis presented by West *et al.* (2010) presented at this meeting and Downing *et al.* (2009).

2 Data and Methods

Patients with colorectal cancer diagnosed between 1998 and 2004 and resident in the Northern and Yorkshire regions were identified from the cancer registry database (n=24,640). Patient age, sex, stage-at-diagnosis (Dukes), and Trust of diagnosis/treatment were extracted and socioeconomic background was represented by the Townsend Index. The binary outcome was dead/ alive at 3 years following diagnosis, since survival at this time has clinical significance.

LCA models are used where subtypes of observations are sought and one wishes to model the uncertainty surrounding observations belonging entirely to one class or another. LCA deals with the uncertainty associated with a limited number of predictor variables when determining subtypes. The LCA models for the colorectal data are multilevel because patients are nested within treatment centres (Trusts). By incorporating discrete latent variables at each level, latent classes at the patient level model uncertainty surrounding patient characteristics and latent classes at the Trust level model variation across Trusts.

To select the number of latent classes at each level, the Bayesian Information Criterion (BIC), the Akaike Information Criterion (AIC) and changes in log-likelihood (LL) were considered, together with model interpretability. All covariates for which there was complete data (age, sex, and SEB) were considered, along with stage at diagnosis (coded A to D for increasing severity and missing coded X); the latter used despite a degree of missing data (13.1%) because stage is key and for categorical variables, it is easy to include a missing category. SMRs were calculated for each Trust and a scaled difference from 'SMR=1' was determined by dividing by the square root of the Trust size. For both the SMRs and the MLLC models, 200 bootstrapped datasets were generated and analysed to yield empirical 95% confidence intervals (CIs).

3 Results

Patients were assigned to two classes of similar size: one with reasonable prognosis (54.3% of cases, of which 63.0% died within three years); one with better prognosis (45.7% of cases, of which 39.3% died within three years). Modally allocating, all patients in patient class one (PC1) diagnosed either at stage B or C died within three years, while in PC2, all patients diagnosed at stage A, B or C survived: stage at diagnosis is an important predictor of survival. The largest Trust class (TC1), with 53.1% of patients, had better prognosis (TC1: 51.3% of patients died within three years; TC2: 53.2% of patients died within three years). Trust ranks are summarised in Table 1; a low ranking value indicates a better survival rate than expected. Figure 1 provides a graphical representation, in order of increasing median probability of belonging to the best survival Trust class by MLLC.

	Median probability of belonging to best survival Trust class	Median Rank (95% CI)	
Trust		ML LC	SMR
1	1.000	1 (1–9.5)	6 (2–11)
2	0.999	3 (1–11)	4 (1–10.5)
3	0.997	4 (1–11)	3 (1–10.5)
4	0.996	4 (1–15)	8 (3–14.5)
5	0.993	5 (1–12.5)	5 (1–13)
6	0.956	8 (2–16)	9 (2–17)
7	0.912	9 (3–17)	5 (1–17)
8	0.908	9 (2–17)	6 (1–18)
9	0.897	9 (3–18)	5 (1–18)
10	0.816	10 (3–17)	8 (1–18)
11	0.575	11 (3.5–18)	11 (3–17)
12	0.476	13 (5.5–18)	12.5 (3–18)
13	0.372	12 (4–18.5)	11.5 (5.5–17)
14	0.359	12 (3–19)	12 (7–17)
15	0.152	14 (5.5–19)	15 (4.5–18)
16	0.070	14 (4–19)	13 (7–18)
17	0.070	15 (7.5–19)	16 (7.5–18)
18	0.003	18 (7–19)	15 (10–18)
19	0.002	18 (13.5–19)	19 (18–19)

TABLE 1. Trust ranks from the MMLC model and the calculation of Trust SMRs.

4 Discussion

By fixing patient class composition and accommodating patient casemix characteristics, the remaining Trust class outcome differences are expected to be variations in Trust performance that depend upon Trust characteristics. The uncertainty surrounding unrecorded patient characteristics was modelled explicitly within the MLLC model – 'fuzzy' matching. We satisfactorily demonstrate the principles of MLLC model casemix adjustment such that differences in the patient pathway of care are modelled explicitly to evaluate organisational features in relation to patient outcomes (e.g. survival, or which treatments are received). This permits hypothesis generation about healthcare delivery and organisational features that warrant intervention to improve patient care. Such studies could inform prospective cluster-randomised trials and are consistent with the MRC framework for the development and evaluation of complex interventions.

An advantage of the MLLC approach is that it does not assume that Trust-level mean outcomes follow a normal distribution, improving 'casemix adjustment'. Trust level covariates may be included and both patient and

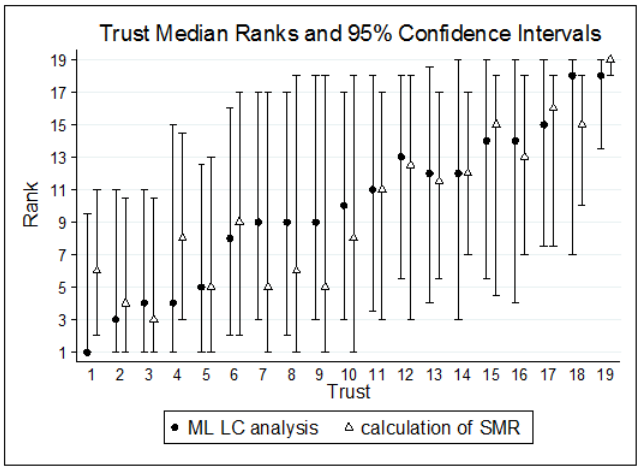


FIGURE 1. Trust median ranks and 95% confidence intervals, ordered by the MLLC analysis.

Trust classes may differ in terms of overall mean outcome and the modelled relationship between the outcome and covariates, capturing additional casemix complexity over and above standard regression models. The principles of this approach extend to other scenarios, e.g. time-to-event analysis, the accommodation of multiple treatment centres, and cross-classified organisational structures. This approach can adjust for variations in the patient pathway (e.g. the delivery of appropriate healthcare), whilst evaluating institutional processes (e.g. delays to treatment) in relation to patient outcomes. This provides a more robust approach to evaluating institutional performance than is currently standard, though further research is needed to demonstrate the utility of modelling different patient pathways and evaluating process differences across institutions.

References

Downing, A., Harrison, W.J., West, R.M., Forman, D., and Gilthorpe, M.S. (2009) Latent class modelling of the association between socioeconomic background and breast cancer survival status at 5 years whilst incorporating stage of disease. *Journal of Epidemiology and Community Health*, in press (epub), PMID: 19692736.

West, R.M., Gilthorpe, M.S., Harrison W.J., Downing, A., and Forman, D. (2010) Modelling an exposure–outcome relationship accommodating potential confounders on the causal path using a latent-class model. *IWSM, Glasgow.*

A Two-Stage Model for Actuarial Run-Off Triangles

Gillian Heller[1], Gerhard Neubauer[2]

[1] Department of Statistics, Macquarie University, Sydney, Australia (corrresponding author: gillian.heller@mq.edu.au)
[2] Institute of Applied Statistics, Joanneum Research, Graz, Austria

Abstract: Insurance claims may occur with a delay of several years and accurate estimates of future claims are of importance for insurance companies. Usually predictions are based on direct modelling of past claim amounts. We propose a method that uses both the number of claims and the claim amounts for prediction. In a first step the number of future claims is predicted from a model for underreported counts. Then the prediction of amounts is accomplished by using the count predictions as offset. The method is applied to real data and it shows good performance.

Keywords: Run-off triangle, insurance claims, underreporting, beta-Poisson, Gamma, regression

1 Introduction

Table 1 shows an extract of insurance accident claims data given by Taylor (2000), classified by the accident year $t = 1, \ldots, T$, and development year $j = 0, \ldots, J$. Here $T = 18$, $J = 10$, $t = 1$ denotes the year 1978, and j is a counter for the number of years after the accident when the claim was settled ("developed"). In addition the first column contains n_t, the number of existing insurance contracts, also known as the exposures. In the actuarial literature matrices such as this are known as run-off triangles, in which cell (t, j) contains either the number of claims, or the total claim amount. The unobserved cells represent future claims for which the insurance company has to provide a reserve. Interest is therefore focussed on prediction of the total unobserved claim amount. In the actuarial literature emphasis has been on direct modelling of the claim amounts. Our approach is to model the number of claims, using methods for underreporting, and to base a model for the total claim amount on expected numbers of claims from this model.

2 Statistical Models

Underreporting Model for Claim Counts

Underreporting is best known from crime data, but may occur with any kind of counts. The estimation of underreporting may be based on a Bernoulli sampling scheme, where $U_i \sim \text{Bernoulli}(\pi)$, $i = 1, \ldots, \lambda$, is an iid sample

TABLE 1. Extract of run-off triangle for number of claims

		j										
n_t	t	0	1	2	3	4	5	6	7	8	9	10
117306	8	573	266	62	12	5	7	6	5	1	0	1
123304	9	582	281	32	27	12	13	6	2	1	0	-
125533	10	545	220	43	18	12	9	5	2	0	-	-
131265	11	509	266	49	22	15	4	8	0	-	-	-
139661	12	589	210	29	17	12	4	9	-	-	-	-
152895	13	564	196	23	12	9	5	-	-	-	-	-
160331	14	607	203	29	9	7	-	-	-	-	-	-
162900	15	674	169	20	12	-	-	-	-	-	-	-
170045	16	619	190	41	-	-	-	-	-	-	-	-
173248	17	660	161	-	-	-	-	-	-	-	-	-
175941	18	660	-	-	-	-	-	-	-	-	-	-

of binary random variables, indicating whether or not an event is reported. Obviously $Y = \sum_i U_i \sim \text{binomial}(\lambda, \pi)$ and the problem is to estimate λ, the number of events that actually occurred and π, the reporting probability. The binomial model is not appropriate when $\text{var}(Y_t) \geq \text{E}(Y_t)$ and for such cases extensions of the binomial model have been developed: the beta-binomial, the negative-binomial and the beta-Poisson model. These models are obtained by randomizing the binomial parameters, and can be characterised by a mean-variance relationship of the form $\text{var}(Y) = \mu\phi$. Further details can be found in Neubauer, Djuraš and Friedl (2010). In the insurance context, the number of claims settled can be seen as underreported cases. If we have λ_t accidents in year t then $\widetilde{y}_{tj}$ claims are settled in the year j after the accident. Hence $y_{tj} = \sum_{k=0}^{j} \widetilde{y}_{tk}$ is the total number of accidents that are settled by j years after accidents in year t. The y_{tj} are unobserved for $j > T - t$. We model these cumulative counts as $\text{E}(Y_{tj}) = \mu_{tj} = \lambda_t \pi_j$, and $\lambda_t = n_t \rho_t$ with $0 \leq \rho_t \leq 1$ the accident risk. Regression addresses λ_t and π_t:

$$\begin{aligned} \log(\lambda_t) &= \log(n_t) + \log(\rho_t) = \log(n_t) + x_t'\beta_x \qquad \text{and} \\ \text{logit}(\pi_j) &= z_j'\beta_z \ . \end{aligned}$$

Mean and Dispersion Model for Claim Amounts

The model for X_{tj}, the total claim amount in year (t, j), is based on $Z_{tj\ell}$, the ℓ-th single claim amount in year (t, j), with $\text{E}(Z_{tj\ell}) = F_1(t)\, F_2(j)$. $F_1(t)$ incorporates the effects of inflation and other year t effects, and $F_2(j)$ is the effect of delay j on the claim size. Later claims tend to be larger because generally these have been litigated. Then $X_{tj} = \sum_{\ell=1}^{\widetilde{y}_{tj}} Z_{tj\ell}$ and hence X_{tj} has a randomly stopped sum distribution (Heller, Stasinopolous and Rigby, 2007). Furthermore $\text{E}(X_{tj}) = \mu_{tj} = F_1(t)\, F_2(j)\, \widetilde{Y}_{tj}$. The Gamma distribution is appropriate for X_{tj} and we use a mean and dispersion parametrization $X_{tj} \sim \text{Gamma}(\mu_{tj}, \sigma_{tj})$, where $\text{E}(X_{tj}) = \mu_{tj}$ and $\text{var}(X_{tj}) = \sigma_{tj}^2 \mu_{tj}^2$. Generalized Additive Models for Location, Scale and Shape (Rigby and

Stasinopoulos, 2005) are well suited for this purpose. We specify the log-linear models

$$\begin{aligned} \log \mu_{tj} &= f_1(t) + f_2(j) + \log \widetilde{Y}_{tj} \qquad \text{and} \\ \log \sigma_{tj} &= g_1(t) + g_2(j), \end{aligned}$$

where $f_1(t) = \log F_1(t)$, $f_2(j) = \log F_2(j)$, and the functions f_1, f_2, g_1 and g_2 may be parametric or smooth.

3 Application to claims data

The overall goal of the data analysis is the reliable prediction of unobserved claim amounts for the period (1986-1995).

Results of the count data analysis

For the count data we use the regression model

$$\begin{aligned} \log(\lambda_t) &= \log(n_t) + \beta_0 + f(t) \qquad \text{and} \\ \text{logit}(\pi_{tj}) &= \alpha_0 + \alpha_1\, j + \alpha_2\, tj \end{aligned}$$

which allows for a time-varying run-off structure in the reporting probabilities. The smooth function $f(t)$ is estimated by a cubic regression spline. The cumulative probabilities π_{tj} are monotone non-decreasing over j. To choose a distributional model we consider the binomial, beta-binomial, negative-binomial and beta-Poisson. The beta-Poisson is the only one which converges and gives reasonable estimates. The interaction term tj has no significant contribution. Figure 1.a shows the cumulative accident data y_{tJ}

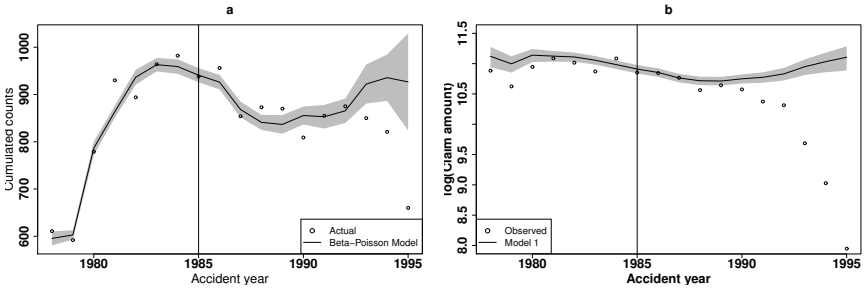


FIGURE 1. a. Cumulative accident counts by accident year and fitted values for the beta-Poisson model; b. Log(claim amount) by accident year, and fitted values for Gamma model 1; with 95% confidence regions. Vertical lines indicate the period with complete observations.

and estimated trend λ_t from the beta-Poisson model. The estimated trend λ_t follows the data until the last but three years. The discrepancy in the

Model	$f_1(t)$	$f_2(j)$	$\log \widetilde{Y}_{tj}$	$g_1(t)$	$g_2(j)$	LL	df	AIC
1	quadr.	quadr.	yes	quadr.	linear	-1247	17	2527
2	quadr.	quadr.	no	quadr.	linear	-1258	9	2535
3	quadr.	quadr.	yes	-	-	-1278	14	2585

TABLE 2. Model selection for Gamma claim amounts model. The degrees of freedom for models 1 and 3 include 8 df for the estimation of $\widetilde{Y}_{tj}$.

last three years is due to the large number of unobserved cells in those years. Note that Y_{tj} are cumulative counts and that the model uses the assumption of independent Y_{tj} in both temporal directions t and j. This strong assumption seems justified as the residuals of our final model are serially uncorrelated in both directions.

Results of the amount data analysis

Using count estimates in a model for claim amounts is contrasted with an equivalent model which does not incorporate the claim numbers information. Results are given in Table 2. The effect of using the count information is seen in the comparison of model 2 with model 1, and clearly improves the model fit. This is an interesting result as actuarial models for claim amounts (e.g. regression models (Taylor 2000); state space models (de Jong 2006)) do not utilise the claim numbers. In addition, specifying the model for σ_{tj} is seen to provide significant improvement, as seen in the comparison of model 1 with model 3. This implies that not only does the mean claim amount vary with accident year and development year, but the dispersion in claim amounts also varies with these factors. Fitted values for model 1 are shown in Figure 1.b. As for the claim numbers, clearly there is a drop-off in observed amounts after 1985, when the period of full observation ends.

References

de Jong, P. (2006). Forecasting runoff triangles. *North American Actuarial Journal.* **10** (2), 28-38.

Heller, G.Z., Stasinopoulos D.M. and Rigby, R.A. (2007). Randomly Stopped Models. *Proceedings of the 22nd IWSM. Barcelona*, 323-328.

Neubauer, G., Djuraš, G. and Friedl, H. (2010). Models for Underreporting: A Bernoulli Sampling Approach for Reported Counts. *Proc. 9th Int. Conf.: Computer Data Analysis and Modeling. Minsk.*

Rigby, R.A. and Stasinopoulos, D.M. (2005). Generalized Additive Models for Location, Scale and Shape. *Applied Statistics*, **54**, 507-554.

Taylor, G.C. (2000). *Loss reserving: an actuarial perspective.* Huebner International Series on Risk, Insurance and Economic Security.

Imputation of household level multivariate discrete data from zonal census data

C.M. Hincksman[1], A.N. Pettitt[1], R.W. Reeves[1]

[1] Queensland University of Technology, GPO Box 2434, Brisbane Queensland 4001, Australia

Abstract: This paper proposes methodology to sample from a disaggregate level model subject to constraints at the aggregate level, using a Metropolis Algorithm. This problem is especially pertinent in survey analysis where only aggregate data is provided to maintain confidentiality. This methodology is applied to a household level model where data are available at the aggregated zone level. The resulting samples of imputed household level data enable the calculation of household level trip number predictions which could not previously be achieved using aggregate zone level census data.

Keywords: Imputation; Constraint; Bayesian; Transportation; Discrete.

1 Introduction

This paper describes a method of imputation for multivariate discrete data that is completely missing and has a constraint on the totals of each variable. Examples of such a problem are found in survey data that has been aggregated in order to maintain confidentiality and non identifiability of records, and yet modelling is desired at the disaggregate level. The difficulty of imputation arises when the marginal distribution conditioned on the constraint is not readily calculable although the unconstrained marginal distributions are available. A simulation study demonstrates the methodology. The work is motivated by a case study in trip production models.

The motivating problem is the imputation of completely missing disaggregate household level census data, given a multi-dimensional marginal model for the household level demographics and constraints on the total of each demographic variable within each zone. The data to build the household level demographic model comes from a 2003/2004 Household Travel Survey in Brisbane, Queensland, Australia. Census data from Brisbane is used to form the zone demographic totals for predictions.

For clarity, we illustrate the problem in terms of allocating categories of persons to households. We consider a two dimensional variable $\mathbf{X}_i = \{X_{i1}, X_{i2}\}$, where $\mathbf{X}_i$ is missing, however the total $\sum_{i=1}^{n} X_{ij} = T_j$ is fixed and known, in each of the two categories, and where the subscript j denotes the categories. For example, we wish to impute the number of adults in each

household i using the multi-dimensional demographic model, and subject to the constraint that the number of persons of each category allocated to each household adds to T_j over the zone.
Conditional independence is used to form the household marginal model

$$p(\mathbf{X}_i|\mathbf{z}_i) = p(X_{i1}|\mathbf{z}_i)p(X_{i2}|X_{i1}, \mathbf{z}_i),$$

where $\mathbf{z}_i$ is a vector of household and zonal explanatory variables. Bounds exist on the range of supported values that X_{ij} can take, such that $X_{ij} \in (a_{ij}, b_{ij})$, and both a_{ij} and b_{ij} may be functions of values of $\mathbf{X}_{i,1:j-1}$. For example, X_{i2} might be restricted to a range $(0, 10)$, with the additional constraint that $X_{i2} \leq X_{i1}$. In this example, $a_{i2} = 0$ and $b_{i2} = \min(X_{i1}, 10)$. The algorithm which produces the imputations is described in Section 2.1. The simulation study which verifies and illustrates the model methodology is in Section 3 and the case study is described in Section 4.

2 Method

2.1 The disaggregation algorithm

This section describes the methodology by which the aggregated data is disaggregated subject to constraints. An initial solution is formed by initially setting the imputation vector X^*_{ij} equal to the conditional mean of each demographic variable, rounded up to the nearest integer, $ceil(E[X_{ij}])$, and then either adding (or subtracting, if appropriate) subjects at random until the required total $\sum X^*_{1:n,j} = T_j$ is achieved.
Once an initial sample is obtained, a Metropolis Algorithm is used to move k subjects to household h_1 from household h_2. At iteration t of the algorithm, denote the number of subjects allocated to household i as $X_i^{(t)}$. Randomly select the pair $h_1 < h_2$ from the values $1, \ldots, n$, and propose to increase or decrease household h_1 by k taken randomly from $1, \ldots, \max_X$ where $\max_X$ is the maximum number of subjects allocatable to each household. When household h_1 is increased, the proposed new household demographic values are

$$\mathbf{X}^* = (X_1^{(t)}, \ldots, X_{h_1}^{(t)} + k, \ldots, X_{h_2}^{(t)} - k, \ldots, X_n^{(t)}),$$

The algorithm requires the calculation of the likelihood of the proposed new sample $p^* = p(\mathbf{X}^*)$, and the current likelihood $p^{(t)} = p(\mathbf{X}^{(t)})$. We then accept or reject this proposed move and with probability $R = \min\left(1, p^*/p^{(t)}\right)$ such that

$$\mathbf{X}^{(t+1)} = \begin{cases} \mathbf{X}^* & \text{with prob. } R, \text{ or} \\ \mathbf{X}^{(t)} & \text{with prob. } 1 - R. \end{cases}$$

Modifications can be made to increase the probability of acceptance. A Gibbs sampling algorithm was also investigated, however we found this Metropolis algorithm faster.

2.2 Measuring between sample variation

We use an idea based upon the variogram. A variance measure for between-sample variation of independent samples $\mathbf{X}_1$ and $\mathbf{X}_2$ is $V = n^{-1}\sum_{i=1}^{n}(X_{1i} - X_{2i})^2$. Consider that there are n_k pairs of sampled imputations k lags apart, i.e. $(\mathbf{X}^{(t)}, \mathbf{X}^{(t+k)})$. The average variance measure for those n_k pairs, $V_k = n_k^{-1}n^{-1}\sum_{t=1}^{n_k}\sum_{i=1}^{n}(X_i^{(t)} - X_i^{(t+k)})^2$, plotted against the lag k, is used to determine if an appropriate number of lags have passed to consider a pair of samples to be independent. The plot should asymptote towards V as k increases and equals V for independent samples.

3 The Simulation Study

A simulation study was performed to demonstrate that the algorithm proposed above samples disaggregate level variables subject to aggregate total constraints. In this simplified simulation study, only one variable needs imputation. It is assumed that there are five persons to be allocated to four households $i = 1, \ldots, 4$ and that the distribution of persons for each household is known. The marginal conditional distribution of X_i^* conditional on the constraint $T = 5$ is calculable because there are only 5^4 possible allocations. The marginal distribution $p(X_i)$ and marginal conditional distributions $p(X_i|T = 5)$ are given in Table 1. A sample is taken after every 100 moves according to the methodology outlined in Section 2.1, until 100000 samples have been recorded, discarding the first 10000 for burn in. These samples are used to estimate the conditional marginal probabilities, also found in Table 1. An acceptable level of precision was achieved in this study.

4 The Case Study

The motivating application for this model is found in transportation planning for Brisbane, Queensland, Australia. Data from a 2003/2004 Household Travel Survey, collected at the disaggregate household level, is used to form a prediction model for household travel demands. However, census data is used to make predictions for future years, and this data has been aggregated to the zone level, containing about 300 households, and hence only zone aggregate totals are available.

In order to predict zonal totals of trips, it has been found that modelling at the household level, rather than aggregated zonal level, gives improved precision of predictions; Kynn (2005). Additionally much modelling has been carried out at the household level where, for example, zero-inflated models have been used; Jang(2005).

There are $C = 176$ zones, indexed by $c = 1, \ldots, 176$. There are n_c households in zone c, indexed by $i = 1, \ldots, n_c$. There are $J = 11$ variables to be

TABLE 1. The Marginal, Marginal Conditional, and observed Marginal Conditional probabilities (OMCP) for the four households in the Simulation Study.

Method	Probabilities	Subject			
		1	2	3	4
Marginals	$P(X=0)$	0.80	0.10	0.50	0.30
	$P(X=1)$	0.10	0.50	0.25	0.65
	$P(X=2)$	0.05	0.25	0.15	0.05
	$P(X=3)$	0.05	0.10	0.10	0.00
	$P(X=4)$	0.00	0.05	0.00	0.00
Marginal	$P(X=0\|T=5)$	0.6785	0.0220	0.2595	0.1745
Conditionals	$P(X=1\|T=5)$	0.1349	0.3812	0.2337	0.7515
	$P(X=2\|T=5)$	0.0940	0.3001	0.2509	0.0740
	$P(X=3\|T=5)$	0.0962	0.1722	0.2559	0.0000
	$P(X=4\|T=5)$	0.0000	0.1235	0.0000	0.0000
OMCP	$P(X=0\|T=5)$	0.6788	0.0215	0.2567	0.1746
	$P(X=1\|T=5)$	0.1351	0.3821	0.2362	0.7512
	$P(X=2\|T=5)$	0.0895	0.3014	0.2527	0.0742
	$P(X=3\|T=5)$	0.0966	0.1731	0.2544	0.0000
	$P(X=4\|T=5)$	0.0000	0.1219	0.0000	0.0000

imputed, indexed by $j = 1, \ldots, 11$, and these are the household size, number of adults, children, primary school aged children, secondary school aged children, tertiary students, dependants A (young children), dependants B (adults), dependants C (elderly), white collar workers and blue collar workers. Three of the variables requiring imputation (X_4, X_7 and X_9) are linear combinations of other variables and are not imputed directly. The variable $X_{c,i,j}$ can take values in the range (a_{cij}, b_{cij}). The imputed variables are needed to make predictions about household level trip counts $Y_1, \ldots, Y_6$ which are blue collar work trips, white collar work trips, primary and secondary school trips, tertiary education trips, shopping trips, and other trips. Within each zone c, the total of the covariates allocated for each variable is $T_{c,j}$.

The distribution of the trip counts strongly favours an even number of trips. This is a result of the tendency of people to make return trips, and has been accounted for through a mixture model of "odd" and "even" trips. In addition, the existing zone level trip count prediction models are linear functions of demographics aggregated to the zone level. As such, they are not able to account for the fact that the per person trip rate is a function of the household size. These features further demonstrate the benefit of modelling at the household level, which requires the imputation of the household demographic variables $\mathbf{X}$.

The household trip count model $p(\mathbf{Y}_i|\mathbf{X}_i)$ is an extension of the work by

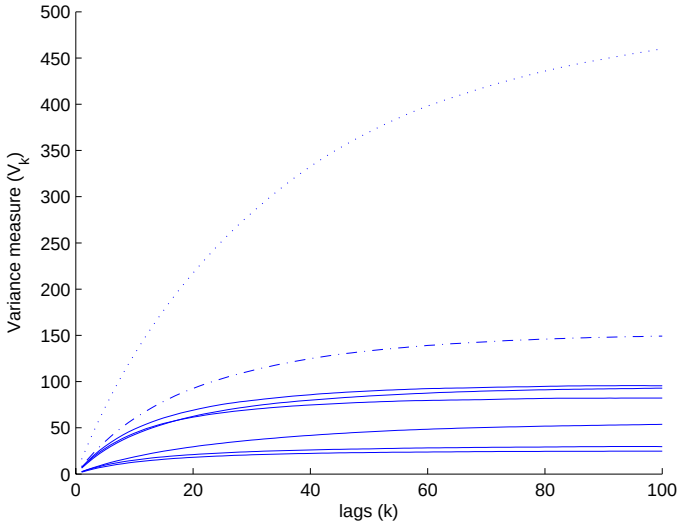


FIGURE 1. This plot shows the between sample variation V_k, described in Section 2.2, reaching its asymptotic limit after approximately $k = 100$ moves, for each of the 8 covariates requiring imputation. Focus is given to the imputation of the household size X_1 (dotted line) and the number of white collar workers X_2 (dot-dash line), and the remaining covariates $X_3, X_5, X_6, X_8, X_{10}, X_{11}$ have their between sample variation plotted in solid lines.

Alfo and Trovato (2004) and the marginal multi-dimensional demographic household models $p(\mathbf{X}_i|\mathbf{T}_i)$ are formed as log-linear models. The details of the fitting and parameter estimates of these models have been omitted.
Figure 1 shows the between sample variation V_k, at k lags, reaching its asymptotic value V after approximately $k = 100$ lags. At this point, samples 100 irerations apart are considered independent. In total, ten samples were imputed using a total of 1000 moves.
Focus now moves on to using these imputed values for the prediction of household level white collar (home based) work trips, which is a function of the number of white collar workers in the household X_2. The prediction of the other trip types is omitted for brevity. The summary statistics (mean, variance) of the observed $\mathbf{X}_2$ from the Household Travel Survey $(0.3600, 0.3511)$, and the imputed samples $\hat{\mathbf{X}}_2$ $(0.3583, 0.4567)$ show a good similarity. The summary statistics of the expected prediction

$$\hat{\mathbf{Y}}_2 = \sum_{y=0}^{\infty} y \sum_{x=0}^{\infty} p(\mathbf{Y}_2 = y|x)p(x|\mathbf{T}_2)$$

using the imputed samples $\hat{\mathbf{X}}_2$ is $(0.5447, 1.301)$. This appears to be different to the summary statistics of the observed trips $\mathbf{Y}_2$ which are $(0.4670, 0.9565)$ and it is hoped that this precision would improve in future work as more of the features and correlation structure at the household level is incorporated into the trip count model $p(\mathbf{Y}_i|\mathbf{X}_i)$.

5 Conclusion

The resulting samples of imputed household person memberships of the different categories enable the calculation of disaggregate household level trip number predictions using zero inflated semi-parametric mixture models extending Alfo and Trovato (2004). This could not be previously achieved using aggregate zone level census data and enables the modelling of the rich structure available only at the household level. A small simulation study verifies the Metropolis algorithm used to produce the disaggregate level imputations.

References

Alfo, M. and Trovato, G. (2004). Semiparametric mixture models for multivariate count data, with application. *Econometrics Journal*, **7**, 426-454.

Jang, T. (2005). Count Data Models for Trip Generation. *Journal of Transportation Engineering*, **131(6)**, 444-450.

Kynn, M. (2005). Analysis of the Trip Production Equations in the BSTM (Brisbane Strategic Transport Model), Technical report, Infrastructure Planning; Planning, Design and Environment Division; Queensland Department of Main Roads.

Bayesian Change Point Detection with Two Wavelet Procedures

S. Huzurbazar[1], A. Chatterjee[2]

[1] Department of Statistics, University of Wyoming Dept 3332, 1000 E. University Avenue, Laramie, WY 82071, USA. email: lata@uwyo.edu
[2] Department of Mathematics, University of Wisconsin-River Falls, 410 S. 3rd Street, River Falls, WI 54022, USA. email: arunendu.chatterjee@uwrf.edu

Keywords: Discrete wavelet transform; lifting; missing data.

1 Introduction

Detection of change points is an important problem in statistics as well as in disciplines such as hydrology and climatology, which use statistical methods for data analysis. In data collected over time, eg. temperature data, change points refer to shifts or changes in the pattern of the data at specific time points. Detection of such changes and estimation of their timing are important as they provide information about the processes generating the observed data.
There are various aspects to the change point problem, namely, detection of a change point, estimation of the time at which the change occured and finally, modelling the data before and after the change. A substantial literature exists on models that combine detection, estimation and modelling using a statistical framework that is either Bayesian or classical. Recent examples include Bayesian linear models for hydrologic data (Seidou et al., 2007) and classical linear models for climatology data (Lu and Lund, 2007). While linear models, especially regression models, are important methods for modelling data, such modelling effort is further complicated when it also includes detection and estimation of change points. We propose that detection and estimation be first explored with a quick algorithm using wavelets.

1.1 Motivation, methods and data

Our work was motivated by water pressure data collected from boreholes on Bench glacier in Alaska. The data were collected at approximately 15 minute intervals, over the course of a year, as part of a larger study to understand aspects of sub-glacial hydrology. Over the course of winter and

early spring, the water pressure records exhibit mainly noise. As the temperatures rise, and the ice in the conduits at the base of the glacier begins to thaw, a change can be found in the water pressure data; thus necessitating a change point detection analysis. Once the change points are detected for all the water pressure records across the boreholes, then the detected change points can be used to assess aspects of the connectivity of the sub-glacial hydrology.

In attempting to detect change points in these data, we explored various methods available in the literature. As modelling the data was not of interest, we concentrated on procedures which only dealt with detection of the change point and estimation of its timing. Wavelets, since they are useful for catching sharp changes and jumps, provide a natural framework for change point detection. The work of Ogden (1996), and Ogden and Lynch (1998) provided a Bayesian procedure using the discrete wavelet transform (DWT). However, it had several shortcomings for our application, as implementing DWT requires no missing data and sample sizes that are powers of two (2^J for J an integer). Consequently, we modified Ogden and Lynch's Bayesian procedure to allow for use of lifting, a recently developed wavelet transform. In this paper, we present an assessment of how well the lifting based procedure works using some simulated data, and then we apply the procedure to streamflow data from the literature, with results for the water pressure data to be presented elsewhere.

2 Discrete Wavelet Transform and Lifting

Wavelets are basis functions that satisfy certain mathematical properties and since the resulting wavelet transforms are localized in time and space, they can be used to detect sharp changes in discontinuous functions. Several good introductions to wavelets exist with Ogden (1997) for statistical applications and Percival and Walden (2000) for time series data. The first generation of wavelets, implementation of which is via the discrete wavelet transform (DWT), maps a data vector $y = (y_1, ..., y_n)$, for $n = 2^J$, to a vector of wavelet coefficients $w = (w_1, ..., w_n)$ via an orthogonal matrix W, so that

$$w = \frac{1}{\sqrt{\mathrm{n}}} W y. \tag{1}$$

Choice of wavelet functions determines W. The first $n-1$ elements of w are indexed as $w_{j,k}$ for $j = 0, ..., J-1, k = 0, ..., 2^j - 1$, and the remaining element is labeled $w_{-1,0}$. The quantity $w_{j,k}$ is called the empirical wavelet coefficient at level j and position $k2^{-j}$. Since higher frequency components occur for larger values of j, detection of change points involves examination of these higher frequency empirical wavelet coefficients in w.

Ogden and Lynch (1998) proposed a Bayesian change point detection procedure using higher frequency empirical wavelet coefficients, to form the

likelihood function, and obtained posterior distributions for the location of the change point in the data. While they simulated data to assess their procedure, in practice, the implementation of the procedure is hampered by the presence of noise in the data, missing data and to some extent, by the sample size not equal to a integer power of 2. We attempted to address these problems by denoising the data, using simple estimates for the missing values as well as for observations in order to augment the sample size to 2^J. These ad hoc methods, which were intended as quick fixes in order to obtain a procedure that worked fast, produced somewhat satisfactory results as documented in Chatterjee (2009).

An improved procedure was obtained by adapting Ogden and Lynch's Bayesian approach to a second generation wavelet transform called lifting which was introduced by Sweldens (1996) with details provided in Sweldens (1997). Lifting overcomes the restrictions of complete data and sample sizes equal to 2^J. In the statistics literature, earlier lifting algorithms were further modified into an adaptive lifting procedure by Nunes et al. (2006) which we use for our change point detection. As with the work by Ogden and Lynch, one aspect of using wavelets for change point detection is the decision of which empirical wavelet coefficients to use in forming the likelihood. While the coefficients have a clear interpretation for DWT, the interpretation as high or low frequency coefficients is less clear for lifting. In our use of lifting, we used a resolution level of 2, which leaves us with (n-2) for the number of 'detail coefficients'. These can be divided somewhat artifically into fine and course levels, but since this classification is somewhat arbitrary, we had to investigate this further. To assess effects of this choice of coefficients and especially its effects in the presence of various conditions present in real data, we conducted a small simulation study.

In what follows, we used the same likelihood as Ogden and Lynch(1998) for the wavelet or lifting coefficients, a normal likelihood which assumes independence of the coefficients, and has as its parameters the location and size of the jump as part of the mean function and the variance. We also used the same priors; namely, a Jeffreys' noninformative prior; details are given in Chatterjee(2009). We assessed the effect of various other priors and found no change, thus, reverting back to the simpler prior.

3 Some Simulation Results

For the various simulations, data were generated from normal distributions with different variances, different levels of autoregression, different degrees of missingness, and different jump sizes at the change point. The interplay between the variance and jump size was investigated, as was the influence of the amount of missing data. For purposes of this short description, we refer to Table 1 for a comparison of DWT and lifting when there is no autocorrelation and no missing data, but as we vary the variance and the

TABLE 1. DWT and Lifting results with no missing data; $n = 2^7$

Size of jump	$\sigma^2 = 0.5$		$\sigma^2 = 1.0$		$\sigma^2 = 1.5$	
	DWT	Lifting	DWT	Lifting	DWT	Lifting
1	54.1%	93.7 %	36.3 %	89.8 %	28.9 %	87.4 %
3	98.0%	93.7 %	91.9 %	91.1 %	84.9 %	90.7 %

TABLE 2. Lifting result for AR(1) data, 4% missing data; $n = 2^7$

Size of jump	$\alpha = 0.2$		$\alpha = 0.6$		$\alpha = 0.85$	
	45% coef	38% coef	45% coef	38% coef	45% coef	38% coef
1	84.2%	23 %	84.2 %	22.9 %	82.5 %	21.9 %
3	89.5%	27.3 %	88.3 %	27.7 %	88.9 %	30 %

jump size. For the harder problem of small jump size, the lifting-based procedure performs much better than DWT. For the easier problem of a larger jump size, the DWT and lifting-based procedures are similar in performance. It should be noted that computationally, DWT is much faster than lifting, a factor to consider as the sample sizes get larger. We next refer to Table 2 for an example illustrating the choice of how many lifting coefficients to use with data in the presence of an AR(1) structure. For this situation, the result is intuitive, more coefficients give better information. In general, this pattern remained constant, with use of 45% of the coefficients giving good results.

4 Some Results for Real Data

For an example of how well these procedures work with real data, and for purposes of comparison, we use annual streamflow data from the St. Lawrence River at Ogdensbourg, New York from the years 1860 to 1950. These data were first analyzed in the change point context by Rasmussen (2001) and subsequently by Seidou et al.(2007), both using variations of Bayesian linear models. A description of the data is given in Rasmussen (2001); here we note that both methods found the mode of the posterior distribution for the time of the change point to be 1891. Figure 1 shows a plot of the data, as well as the posterior distribution, using 45% of the lifting coefficients, of the time of the change point. The multimodal posterior has its global mode at 1891. In this case, using DWT also gave a similar result, though since the original data had 90 observations, so we augmented the size of the data to n=128 or 2^7 using an ad hoc procedure of adding values

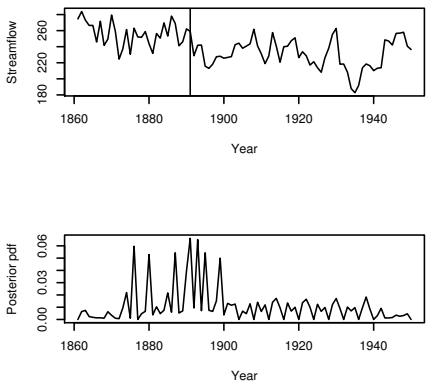


FIGURE 1. St Laurence Streamflow: Original data with changepoint marked and Posterior pdf)

generated using a $N(\tilde{Y}, \hat{\sigma}^2)$, where $\tilde{Y}$ is the median of the data and $\hat{\sigma}^2$ is the estimated variance.
For the water pressure data, we obtained similar results, which we do not present here due to space limitations. In that application, we do not know the real timings of the change points, though the timings obtained make sense given the timing of melting.

5 Discussion

This work was motivated by a very real problem of identifying change points in real data. The literature on change point detection is vast, and has many procedures, some of which are fairly complicated, and most of which rely on modelling of the data. In our experience, modelling the water pressure data is extremely complicated, and such modelling was not the goal of the glaciology research. The seemingly simpler question had to do with detection and estimation of the timing of the change points. Wavelets provide a quick, fairly non-parametric approach for such detection. We modified an existing procedure based on the discrete wavelet transform, and also extended the procedure for use with a newer wavelet transform. Both gave good results, with the DWT procedure being somewhat ad hoc in its application in the presence of missing and noisy data, but being faster computationally. We presented results for a dataset which has been used in the literature. Results for the water pressure data and other examples from the literature are in Chatterjee (2009) and are being included in

manuscripts currently in progress.

Acknowledgments: Special Thanks, for providing the water pressure data, to Dr. Neil Humphrey, Dept of Geology & Geophysics, University of Wyoming, USA

References

Chatterjee, A. (2009). Detection of Change Points using Wavelet Analysis. Unpublished Ph.D. thesis at the University of Wyoming.

Heaton, T. J., and Silverman, B. W. (2008). A wavelet or lifting-scheme-based imputation method. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **70**, 567-587.

Lu, Q. and Lund, R.B. (2007). Simple linear regression with multiple level shifts. *Canadian Journal of Statistics*, **35 (3)**, 447-458.

Nunes, M., Knight, M., and Nason, G. P. (2006). "Adaptive lifting for non-parametric regression", *Statist. Comput.* **16**, 143-159.

Ogden, R. Todd (1996). Wavelets in Bayesian change-point analysis. *ASA Proceedings of the Section on Bayesian Statistical Science*, 164-169, American Statistical Association, Alexandria, VA.

Ogden, R. Todd (1997). *Essential Wavelets for Statistical Applications and Data Analysis*, Birkhauser, Boston.

Ogden, T., and Lynch, J. D. (1998). Bayesian analysis of Change-point models. in *Bayesian Inference in Wavelet Based Models*, P. Muller and B. Vidokovic eds., Springer-Verlag: New York.

Percival, Donald B., and Walden, Andrew T. (2000). *Wavelet methods for time series analysis* , Cambridge University Press: Cambridge, U.K.

Rasmussen, P. (2001). Bayesian estimation of change points using the general linear model. *Water Resources Research*, **37**(11), 2723-2731.

Seidou, O., and Ouarda, T. B. M. J. (2007). Recursion-based multiple change-point detection in multiple linear regression and application to river streamflows. *Water Resources Research*, **43**, W07404, doi:10.1029/2006WR005021.

Sweldens, W. (1996) Wavelets and the lifting scheme: A 5 minute tour. *Z. Angew. Math. Mech.*, **76**, Suppl. 2, 41-44.

Sweldens, W. (1997). The lifting scheme: A construction of second generation wavelets. *Siam J. Math. Anal*, **29**, No. 2, 511-546.

Robust Survival Trees Based on Node Resampling

Alberto Alvarez Iglesias[1], John Newell[1,2], John Hinde[1], Liam Glynn[3]

[1] School of Mathematics, Statistics and Applied Mathematics, NUI, Galway, Ireland. (a.alvareziglesias1@nuigalway.ie)
[2] HRB Clinical Research Facility, NUI, Galway, Ireland.
[3] Department of General Practice, National University of Ireland, Galway, Ireland

Keywords: Recursive partitioning; Survival Trees; Random Survival Forest.

1 Introduction

Recursive partitioning methods are a popular non-parametric alternative to the classical parametric and semi-parametric models used in regression, classification and survival problems. They have been recognised as a useful modelling tool as they produce a model that is very easy to interpret. The beauty of these methods is their simplicity and the relative ease in which the results of the analysis can be explained to a person with a non statistical background. From the statistical point of view, however, criticisms arise from the lack of statistical tests and problems due to overfitting. In general, the aim of any statistical model is to explain the complicated structure evident in a large number of explanatory variables and the response, in the simplest way possible, and to use the model to predict the outcome of interest when new observations are considered. Trees are an excellent way to describe the structure of the learning data but their predictive power can be disappointing. In the last decade, many efforts have been made to overcome this problem. These methods are generally known as "ensemble methods" and they use a set of trees, created by bootstrapping the original data, in order to improve predictability. The price to be paid, however, is the absence of a singular tree.
In this work, data on 1586 patients with cardiovascular disease will be analyzed and the results of different methods for growing survival trees will be compared. The event of interest was a cardiovascular composite endpoint, which included death from a cardiovascular cause or any of the cardiovascular events of myocardial infarction (MI), heart failure (HF), peripheral vascular disease (PVD) and stroke. Seventeen explanatory variables were considered for development of a prognostic model.

In an effort to combine the simplicity of a single tree and some of the ideas in which ensemble methods are based, a new method will be proposed which is a composite of both procedures where resampling is made at node level with a single tree as the output.

2 Survival trees

Survival data are special due to the presence of censoring. Many attempts have been made to use recursive partitioning to create survival trees. Davis and Anderson (1989) proposed a splitting criterion based on likelihood assuming the exponential model. Leblanc and Crowley (1992) proposed a more general method based on the assumption of the proportional hazards model.

Segal (1988) created an algorithm based on the log-rank statistic as a measure of dissimilarity in the splitting process. Segal argued that tests based on measures of dissimilarity between nodes can tell more about the important prognostic factors associated with the survival times than within node homogeneity. Ensemble methods have been developed to create a forest of survival trees by Breiman (2003), and more recently by Ishwaran (2008).

Several packages in R have implemented methods to grow survival trees. Figure 1 shows a tree grown using *rpart* based on the deviance residuals proposed by Leblanc and Crowley (1992).

More recently, a new package named *party* has been developed using conditional inference procedures where splits are based on a p-value with a Bonferroni adjustment used to ensure a desired family wise error rate. Figure 2 shows such a tree based on the log-rank statistic as splitting criterion. As a summary for the terminal nodes, Kaplan-Meier estimates of the survivor function are given.

Finally, Figure 3 shows the results of an ensemble of survival trees generated by the package *randomSurvivalForest*. The method consist of a forest of survival trees by bootstrapping samples of the original data, using, in this example, the log-rank test as splitting criterion. In this procedure, only a set of randomly selected explanatory variables are considered at each step. The resulting output from this approach is an ensemble cumulative hazard function which average the Nelson-Aalen estimators of the cumulative hazard functions of each one of the trees in the forest. To estimate prediction error Harrell's concordance index is considered (Harrell et al. 1996). The plot presents the prediction errors and a measure of variable importance (see Ishwaran 2008).

When comparing the three methods, Age appears as the most useful prognostic indicator, with a splitting value of 77 years, in two of the trees. The random survival forest, however, suggests *Previous PVD* as the most important variable despite it not appearing in the *rpart* and *party* output.

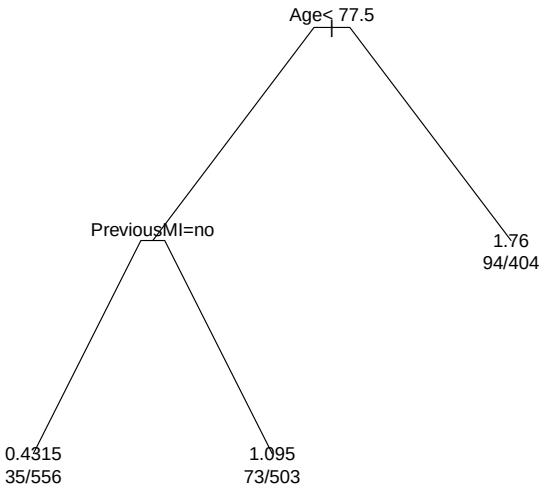


FIGURE 1. *r-part* survival tree based on deviance residuals under the assumption of proportional hazards. The summaries in the terminal nodes represent relative risks and number of deaths/number of cases.

3 Robust Survival Trees

The approach considered in this paper is to use resampling at the node level where the explanatory variable for each split is chosen as the candidate that most often appears as the best predictor for each one of the bootstrap replicates. Log-ranks will be considered as splitting criterion although other criteria can be acomodated. As an example of this technique, Figure 4 shows the results for the primary split. *Age* is the best candidate and *Cholesterol*, *Diabetes* and *Previous HF* could be considered as surrogate splits.
Once the explanatory variable is chosen for the split, bootstrap estimates of the cutpoints and standard errors are available. Figure 5 shows the distribution of the cutpoints for *Age* at the primary node. The mean and standard deviation are 78.01 and 3.20 respectively and a 95% bootstrap interval for the cut points is $(71.61, 83.47)$.

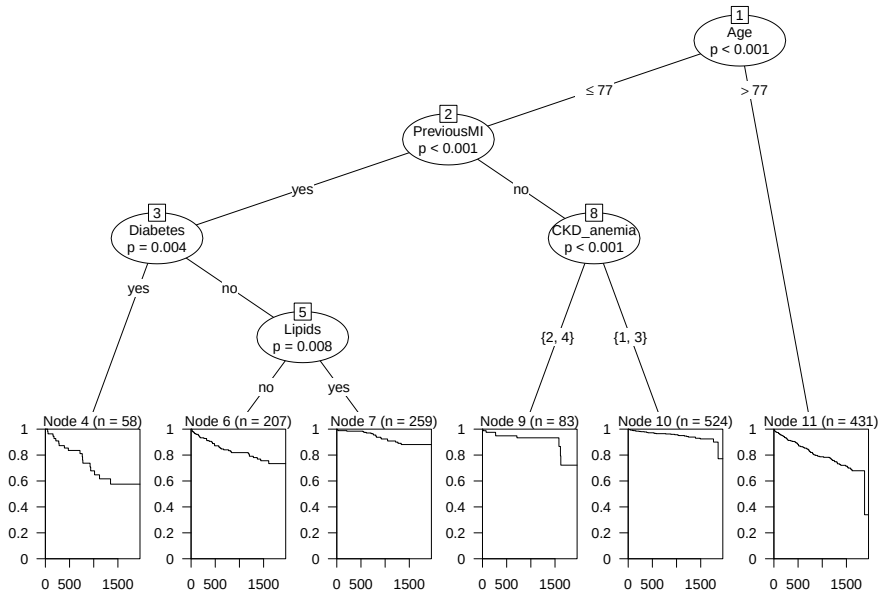


FIGURE 2. *party* survival tree based on log-rank test statistic as a measure of dissimilarity between nodes. In the terminal nodes, Kaplan-Meier estimates of the survivor functions are given.

After the tree has been grown and pruned, prediction errors will be assessed using Harrell's concordance index (Harrell 1996). The results will be compared to those produced by the survival tree methods mentioned previously in terms of predictability and ability to capture and summarise the complex structure of the data.

Acknowledgments: The first author is grateful to IRCSET (the Irish Research Council for Science, Engineering and Technology) for their continued funding of postgraduate students.

References

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984), *Classification and Regression Trees*. Wadsworth, Belmont, CA.

Breiman, L. (2003). Manual setting up, using and understanding random forests V4.0. Available at ftp://ftp.stat.berkeley.edu/pub/users/breiman/Using_random_forests_v4.0.pdf

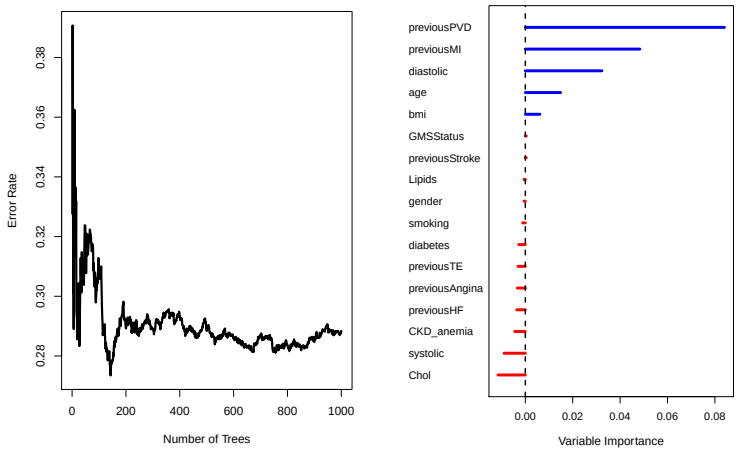


FIGURE 3. Random survival forest output. The left panel displays a measure of the prediction error while the right panel displays a measure of variable importance.

Davis, R.B. and Anderson, J.R. (1989). Exponential survival trees. *Statist. Med.* **8**, 947-961.

Harrell, F.E., Lee K.L., Mark, D.B. (1996). Tutorial in biostatistics. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15**, 361-387.

Ishwaran, H., Kogalur, U.B., Blackstone, E.H. and Lauer, M.S. (2008). Random survival forest. *The Annals of Applied Statistics* **2**, 841-860.

Leblanc, M. and Crowley, J. (1992). Relative risk trees for censored survival data, *Biometrics*, **48**, 411-425.

Segal, M. (1988). Regression trees for censored data. *Biometrics* **44**, 35-48.

Zhang, H.P. (1995). Spliting criteria in Survival Trees, in *Proceedings of the 10-th IWSM, Innsbruck Austria, July 1995*, 305-314.

Zhang, H.P. and Singer, B. (1999). *Recursive Partitioning in the Health Science*, Springer.

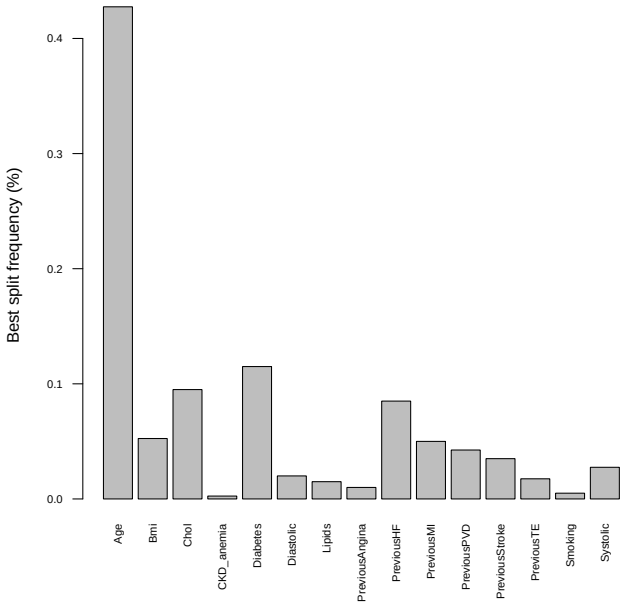


FIGURE 4. Bar chart of the candidate predictors. Each bar represents the proportion of times the explanatory variable has been chosen as the best predictor.

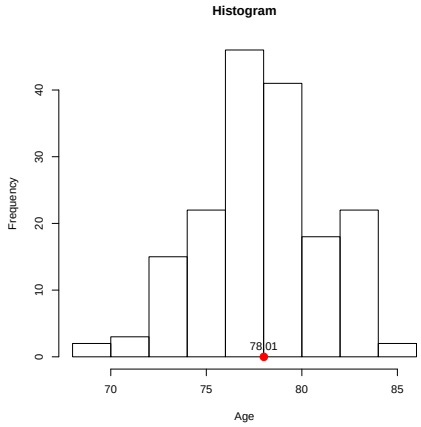


FIGURE 5. Distribution of the cutpoints (with mean highlighted).

Modeling internet congestion: lessons from Brownian motion

S.Ravi Jagannathan[1], K.M. Matawie[1],[2]

[1] School of Computing and Mathematics, University of Western Sydney, PO Box 10, Kingswood NSW 2747, Australia. email: s.jagannathan@uws.edu.au
[2] k.matawie@uws.edu.au

Abstract: The congestion control method Sierra was developed and explicated elsewhere in several papers, and has been shown to have close implications for the statistical modeling, suggesting and using Brownian motion formulation. The latter topic was reported first by Robert Brown, in 1828. The problem was debated extensively by mathematical physicists for several decades after Brown; over the years a definitive treatment of the subject was given by Wiener(1976) formalized in the terms of contemporary mathematics. Again, this work is recapitulated here from our perspective of modeling Black Box congestion control protocols.

Keywords: Congestion Control, Brownian Motion, Sierra, Wiener Process

1 The structure of Sierra

Sierra is a simple, yet novel, Black Box algorithm that uses three basic parameters to optimize network usage. Sierra has been further explicated elsewhere Jagannathan and Matawie(2009) and in other work in preparation. " RTT (Round Trip Times) " Ingress Rates (Rin) " Egress Rates (Rout) No exponential smoothing is performed. Briefly, the control equations for SSS (Sierra Quick Start) and SCAM (Sierra Congestion Avoidance Mechanism) are as follows:
SSS:
$Rin/ = 2$
$UntilRin \leq Rout$
SCAM:
$Rin++$
$IfRin \leq Rout$
$Rin-$
$IfRin > Rout$
Rin and Rout pertain to the transmission rates at Ingress and Egress. Egress rates are continually calculated at the receiver and relayed back to the sender, using control packets. We use this intuitively appealing approach to determine the average steady state throughput for Sierra.

2 The proposed model

The intent of this poster is to set out the statistical model for dealing with the quantitative determination of the throughput performance of Sierra, with a view to the tuning of the parameters to optimize its performance. Sierra is closely linked to Brownian motion. This was established in Jagannathan and Matawie(2009. We give a summary of the stochastic process terminology in this connection. To quote from the literature "it must be assumed that each single particle executes a movement independent of the movement of all other particles; the movements of one and the same particle after different intervals of time must be considered as mutually independent processes, so long as we think of these intervals of time as being chosen not too small. We will calculate the distribution of the particles at time $t = \Delta$ from the distribution at the time t." Our intent is to try to introduce various classical techniques from the stochastic calculus to model the performance characteristics of Sierra, and establish its superiority in this connection.
Consider a particle sitting initially at the origin of R^2 (Cartesian 2-space) located initially at the origin. The position of the particle at time t is a random variable (rather a random 2-vector)

$$X(t) = (X_1(t), X_2(t)) \rightarrow RandomVariable$$

We need to determine $P(X(t))\epsilon B_\Delta$ where $B_\Delta \leq R^2$ such that Measure $B_\Delta \geq 0$. It can be demonstrated that

$$P(X(t)\epsilon B_\Delta = \int K(X,t)d^2x$$

Where $K(X,t) = (1/\sqrt{(}2\pi t))exp(-x2/2t)$ All integrals being in $L^2(R)$.The stochastic process set up above has the following property:

1. $X(0), X(t_1) - X(t_0), X(t_2) - X(t_1), , X(t_n) - X(t_{n-1)})$ are independent
2. $X(t) - X(s)$ has the same distribution as $X(t+h) - X(s+h)$, for all $s, t, h \geq 0$ and $s < t$
3. $X(t) - X(s)$ is Normally distributed
4. $E[X(t)] = 0$

We see from the discussion above, that the Wiener process is a correct fit for the standard Brownian movement problem. We will need to further customize and explicate this process and its implications for the purposes of congestion control. Considerable tailoring will be required Durtt (1998). There will also be a significant interface with Control Systems Theory Kalman et al(1969).

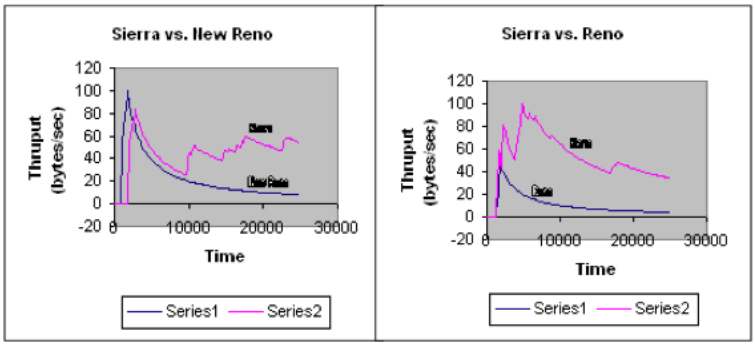


FIGURE 1.

3 Data gathering and simulation

In the figures above we present some information gathered from our recent simulation exercises in software.
As will be noted, Sierra appears to provide more throughput than say Reno or New Reno, while at once being fair to them. We are in the process of collecting real data to offer another mechanism in support of what we are saying in this paper.

4 Current and future work

In other work that is in preparation, we reengineer the congestion control algorithm Sierra along the lines of a specially customized Wiener process technology and its attendant implications. Expressions for throughput performance will be determined and set out, in stochastic terms. There will be opportunity for the tuning of Sierra parameters, with a view to optimize the performance of the protocol. Further simulation will also be carried out in software, to experimentally compare the optimized performance levels of Sierra vis--vis its other competitors.

References

Durett, R., 1998. Stochastic Calculus: A Practical Introduction, CRC Press.

Jagannathan, S. Ravi and Matawie, K. M., 2009. Modeling and Simulation of Congestion Control Algorithms, Business Review Cambridge, vol. 13, No. 2.

Kalman, R. E., et al.1969. Topics in Mathematical Systems Theory,McGraw Hill.

Wiener, N., 1976. Collected Works, vol.1, MIT Press.

A longitudinal model for multiple diagnostic tests: Bovine digital dermatitis

Geoff Jones[1], Wes Johnson[2], Daan Vink[3], Nigel French[3]

[1] Fundamental Sciences (Statistics), Massey University, Palmerston North, New Zealand
[2] Department of Statistics, University of California, Irvine, U.S.A.
[3] EpiCentre, Institute of Veterinary, Animal & Biomedical Sciences, Massey University, New Zealand

Abstract: For many diseases the infection status of individuals cannot be observed directly, but can only be inferred from biomarkers that are subject to measurement error. Diagnosis based on observed symptoms can itself be regarded as an imperfect test of infection status. The temporal relationship between infection and disease may be complex, especially for recurrent diseases where individuals can experience multiple bouts of infection. Given repeated measures of a biomarker for infection and apparent disease status of a number of individuals at multiple time points, together with relevant covariates, we propose and estimate a model in which the unobserved infection status is a correlated latent process. This model can be used to investigate the temporal dynamics of infection, and to evaluate the usefulness of the biomarker for monitoring purposes. Our work is motivated and illustrated by a longitudinal study of Bovine Digital Dermatitis on commercial dairy farms in Cheshire, UK.

Keywords: Bayesian statistics; Diagnostic Test; MCMC; Sensitivity; Specificity.

1 Introduction

Statistical analyses of diagnostic test data for screening or monitoring populations for a disease often rely on a cross-sectional study only. If longitudinal data are available giving the results of testing a number of individuals at multiple time points, a much richer class of modelling opportunities becomes available, including an investigation of the natural history of the disease.

1.1 Relevant literature

The use of imperfect diagnostic tests in a longitudinal study to infer the natural history or temporal characteristics of a disease has been considered by a number of authors. Most have been for progressive, rather than recurrent diseases, most have considered only one diagnostic test, possibly without

error, and all have considered covariate effects in a very limited way. Chen *et al.* (1996) model the progression in breast cancer tumours using a three-state continuous-time Markov chain, based on regular mammography scans. Satten and Longini (1996) estimate a seven-state continuous-time Markov chain for progression of human immunodeficiency virus in HIV-infected men, based on CD4 cell count – a continuous biomarker observed with error. Craig *et al.* (1999) employ a discrete-time non-homogenous Markov model for transitions between five levels of diabetic retinopathy, defined from observed values of retinopathy level and visual acuity. Jackson and Sharples (2002) model the progression of bronchilitis obliterans syndrome in lung transplant patients as a continuous-time Markov process with three states. A continuous biomarker, forced expiratory volume (FEV) and a clinical judgement of the state of each patient are available at irregularly-spaced time points. Inoue *et al.* (2008) use a continuous biomarker for prostate cancer, PSA, to classify individuals into healthy, localized or metastatic states, observed with error at the time of clinical diagnosis. Norris *et al.* (2009) consider a continuous-time model for Johne's disease in cattle with three progressive states and two repeated tests, a continuous serology score and a binary faecal culture test.

In this paper we consider a two-state discrete-time model for recurrent diseases for which a longitudinal study provides repeated measures of a number of non-gold standard diagnostic tests, binary or continuous, with possible fixed and random covariate effects on both the test characteristics and the transition rates.

1.2 Illustration: Bovine digital dermatitis

Bovine digital dermatitis (BDD) is a disease causing lesions on the feet of infected cows. A recently developed enzyme-linked immunosorbent assay (ELISA), that measures the level of serum antibodies, offers an alternative to the standard, but time– and labour–intensive, diagnosis by foot inspection. The performance of this ELISA and its association with foot inspection were examined in a cross-sectional study reported in Jones *et al.* (2009), who also found significant covariate effects for age and foot hygiene score (FHS).

Very little scientific work has been done on the investigation of temporal trends of BDD. It is important to understand the temporal dynamics of the disease as this has consequences for control and intervention strategies.A prospective cohort study was carried out on four farms in Cheshire, UK in 2004–05. Observations on each sampled cow were made at approximately equal intervals averaging about three weeks. At each sampling point, blood samples were obtained for serological analysis, foot hygiene was scored and feet were inspected for lesions. Our dataset has observations on 119 cows, with on average 13 observations per cow. The analysis of Jones *et al.* (2009) suggests that a log transform of the raw ELISA scores achieves approximate

normality, that age may have an approximately quadratic effect and that FHS should be log-transformed. We define

$$\begin{aligned} S_{ij} &= \text{log of ELISA score for cow } i \text{ at observation time } j \\ T_{ij} &= \text{presence of lesions (0/1)} \\ x_{1,ij} &= \text{mean-centred log(FHS) – high score implies poor hygiene} \\ x_{2,ij} &= \text{mean-centred age in years} \\ x_{3,ij} &= x_{2,ij}^2 \end{aligned}$$

Seasonality in a plot of the clinical prevalence, as defined by the presence of lesions, across all farms during the study period suggests adding two further covariates

$$\begin{aligned} x_{4,ij} &= \sin(2\pi d_j/365.25) \\ x_{5,ij} &= \cos(2\pi d_j/365.25) \end{aligned}$$

where d_j is the number of days since the start of the study. Possible random effects could be associated with individual cows, say Cow_i and with cows in the same management group at the same time, MGT_g.

2 Model

Our model consists of two parts, one relating current, but unobservable, infection status to previous infection status and covariates, and another relating the observable test results to current infection status, with possible covariate effects. For reasons of parsimony and identifiability we avoid having the same covariate or random effect structure in different equations. Preliminary univariate analyses of lesion status and serology score showed evidence of seasonality in lesions but not serology score, whereas age and FHS affect both serology and lesion status in similar ways. A parsimonious way to model this is to assume that age and FHS are affecting the unobserved rate of infection, whereas seasonality affects not infection status but the probability of lesions given infection. Specificity of lesions, *i.e.* the probability of falsely observing lesions when there is no infection, is treated as a constant.

We model infection status I_{ij} as Bernoulli($\pi_{ij}^{(I)}$) with

$$\text{logit}\pi_{ij}^{(I)} = \alpha_1 + x_{1,ij}\beta_1 + x_{2,ij}\beta_2 + x_{3,ij}\beta_3 + \gamma_1 I_{i(j-1)} + \gamma_2 \pi_{g(i,(j-1))}^{(g)}(1 - I_{i(j-1)}) \quad (1)$$

where $\pi_g^{(g)}$ is the prevalence of infection in management group g. Here γ_1 measures the tendency for an infected cow to stay infected, and γ_2 the tendency for uninfected cows to pick up an infection from others in the same group. Group prevalences can be estimated from the imputed I_{ij}s of the cows in each group.

Conditional on infection status, lesion status is Bernoulli($\pi_{ij}^{(T)}$) with

$$\text{logit}\pi_{ij}^{(T)} = I_{ij}Se_{ij} + (1 - I_{ij})Sp \tag{2}$$

where

$$Se_{ij} = \alpha^{(Se)} + x_{4,ij}\beta_4 + x_{5,ij}\beta_5 \tag{3}$$

This incorporates seasonality into the sensitivity of lesion diagnosis as a test for infection status. The specificity of this test is assumed constant. The continuous–valued serology score is assumed, after log transformation, to be normally distributed conditional on infection status

$$S_{ij} = \mu + \Delta_i I_{ij} + \text{Cow}_i + \epsilon_{ij} \tag{4}$$

where ϵ_{ij} is an AR(1) process with autoregressive parameter ψ to allow for smooth variation in the score and Δ_i is the cow-specific increase in serology score with infection, assumed normal with mean μ_Δ and precision τ_Δ.
For model fitting and inference we use a Bayesian approach, first specifying priors on the model parameters and using Markov chain Monte Carlo techniques to sample from the joint posterior distribution. Where prior information is available we use informative priors. In particular, we incorporated prior information on the specificity of lesions Sp, the average sensitivity of lesions $\exp(\alpha^{(Se)})/[1 + \exp(\alpha^{(Se)})]$ and the distributions of serology scores μ, μ_Δ. For all other parameters we specify vague priors. Note that normal(0,1) priors are vague priors on the logistic scale. A list of model parameters and priors is given in Table 1.

3 Results

The model as specified above was programmed and run in WinBUGS (Lunn *et al.*, 2000). Initially three separate chains were run and the Gelman-Rubin convergence statistic was used to assess a suitable burn-in period. The first 10 000 samples were discarded and the chains were then thinned to every 50th sample as there was strong autocorrelation in some variables. We then took 4000 further samples in each chain and combined them to produce our posterior estimates. Fitting this model in WinBugs we obtained an unequivocally negative value for γ_2, suggesting contrary to expectations that uninfected cows were more likely to stay uninfected if in a high-prevalence group. An epidemiological explanation was put forward that some cows develop immunity to BDD, and the maintenance of this immunity requires that it be constantly challenged by exposure. It would be expected however that this immunity would only develop in older cows, say > 3 years. To test this, we added to (1) an interaction with age $\gamma_3 x_{2,ij}\pi_{g(i,(j-1))}^{(g)}(1 - I_{i(j-1)})$. This gives our final model, the results for which are shown in Table 1. If we examine the estimated value of $\gamma_2 + \gamma_3 x_2$ for a range of ages, we find

TABLE 1. Estimates of final model parameters, corresponding prior distributions and posterior summary statistics. Point estimates are the posterior medians.

parameter	prior	estimate	2.50%	97.50%	MC error
α_0	normal(0, 1)	0.033	-0.462	0.547	0.0033
α_1	normal(0, 1)	-1.632	-2.270	-0.996	0.0037
β_1	normal(0, 1)	1.000	0.091	1.930	0.0062
β_2	normal(0, 1)	0.333	0.185	0.493	0.0014
β_3	normal(0, 1)	-0.052	-0.095	-0.010	0.0002
γ_1	normal(0, 1)	5.007	4.258	5.803	0.0088
γ_2	normal(0, 1)	-2.052	-3.435	-0.780	0.0092
γ_3	normal(0, 1)	-1.789	-2.520	-1.028	0.0097
μ	normal(0, 2)	1.069	0.982	1.157	0.0010
μ_Δ	normal(1, 2)	0.439	0.332	0.548	0.0021
τ_c	gamma(.001, .001)	5.261	3.940	6.911	0.0141
τ_Δ	gamma(.001, .001)	11.48	7.017	28.25	0.2921
ψ	beta(1,1)	0.461	0.408	0.518	0.0006
τ_s	gamma(.001, .001)	130.5	117.0	143.4	0.1908
α^{Se}	logit[beta(2.75, 1.9)]	0.711	0.667	0.728	0.0001
β_4	normal(0, 1)	-0.220	-0.429	-0.009	0.0018
β_5	normal(0, 1)	0.399	0.175	0.624	0.0019
Sp	beta(42.6, 5.62)	0.934	0.911	0.954	0.0003

that it is positive for ages 3 and below, negative for ages 4 and above. This supports the theory that older cows develop resistance that needs regular exposure to be maintained.

Predictive probabilities of infection for each cow at each observation time are easily obtained from the MCMC fitting by monitoring and summarising each I_{ij}. Figure 1 shows the inferred infection status for one cow together with its serology scores and lesion status.

References

Chen, H.H., Duffy, S.W., and Tabar, L. (1996). A Markov chain method to estimate the tumour progression rate from preclinical to clinical phase, sensitivity and positive predictive value for mammography in breast cancer screening. *The Statistician*, **45**, 307-317.,

Craig, B.A., Fryback, D.G., Klein, R., and Klein, B.E. (1999). A bayesian approcah to modelling the natural history of a chronic condition from observations with interventions. *Statistics in Medicine*, **18**, 1335-1371.,

Inoue, L.Y.T., Etzioni, R., Morrell, C., and Müller, P. (2008). Modelling disease progression with longitudinal biomarkers. *Journal of the American Statistical Association*, **103**, 259–270.,

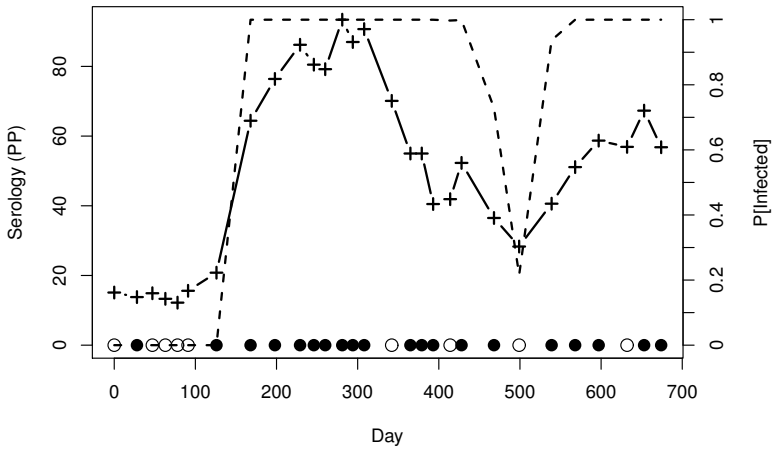


FIGURE 1. Serolgy scores (crosses), lesion status (circles: • = lesions) and predictive probability of infection (dotted line) for one cow.

Jackson, C.H., and Sharples, L.D. (2002). Hidden Markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Statistics in Medicine*, **21**, 113-128.,

Jones G., Johnson W.O., and Vink W.D. (2009). Evaluating a continuous biomarker for infection by using observed disease status with covariate effects on disease. *Applied Statistics* **58**, 705-717.,

Norris, M., Johnson, W.O., and Gardner, I.A. (2009). Modeling bivariate longitudinal diagnostic outcome data in the absence of a gold standard. *Statistics and its Interface*, **2**, 171-185.,

Satten, G.A., and Longini, I.M. (2008). Bayesian latent class models with conditionally dependent diagnostic tests: A case study. *Statistics in Medicine*, **27**, 4469-4488.,

Com-Poisson versus Negative-binomial Models for Over-dispersed Longitudinal Count Data

V. Jowaheer[1], N. Mamode Khan[2]

[1] Department of Mathematics, University of Mauritius, Reduit, Mauritius.

Abstract: Poisson distribution is either mixed with another distribution or generalized to model the over-dispersion present in count responses. Negative-binomial distribution and Com-Poisson distribution are two such Poisson-mixed and generalised Poisson distributions respectively that have been used to model the effects of covariates on the responses in regression set-ups. In this paper, we discuss the models based on these two distributions to analyze over-dispersed count responses arising in longitudinal studies. A simulation study is carried out to compare the performance of both models.

Keywords: Longitudinal data; Count responses; Over-dispersion; Com-Poisson and Negative-binomial Distributions.

1 Introduction

There are in general two classes of distributions which account for over-dispersion relative to Poisson distribution. These are Poisson mixed distributions (PMDs) and Generalized Poisson distributions (GPDs). PMDs include well-known Poisson-gamma (negative-binomial) and Poisson-lognormal distributions. GPDs are obtained by applying normalizing weights to the Poisson distribution and include the exponentially weighted Poisson distributions and the Com-Poisson distribution [Shmueli et al. (2005)] among others. In order to study the effect of covariates on the over-dispersed count responses, models have been designed based on these two classes of distributions. Models based on PMDs are extensively studied by several researchers. It is remarked that negative-binomial model performs better than log-normal mixed Poisson model for a wide range of over-dispersion parameter [Jowaheer (2006)]. Modelling based on GPDs is not very common. However, for some applications, one can refer to Famoye et. al. (2004), Guikema & Gofelt (2008). Recently, Jowaheer & MamodeKhan (2009) have developed a model based on Com-Poisson distribution to analyse the over-dispersed count responses in cross-sectional studies. In this paper, we discuss the models based on Com-Poisson as well as negative-binomial distribution to analyze over-dispersed count responses arising in longitudinal

studies. The estimates of the regression parameters under both models are obtained using joint generalized quasi-likelihood ($JGQL$) estimation method. A simulation study is carried out to compare the performance of both models.

Let y_{it} be the count response of the i^{th} individual at time t ($i = 1, 2, \ldots, I; t = 1, 2, \ldots, T$). Let x_{it} be the p dimensional vector of covariates corresponding to y_{it}. Let β be the p dimensional vector of regression parameters associated with covariate x_{it}. The count responses are over-dispersed and the T responses of the i^{th} individual tend to be correlated and assumed to follow Gaussian type AR(1) autocorrelation structure, as suggested by Jowaheer & Sutradhar (2002), with correlation parameter ρ ($0 < \rho < 1$).

2 Longitudinal Models

2.1 Com-Poisson Longitudinal Model

The Com-Poisson (CP) regression model is formulated as

$$f(y_{it}) = \frac{\lambda_{it}^{y_{it}}}{(y_{it}!)^{\nu}} \frac{1}{Z(\lambda_{it}, \nu)};\ Z(\lambda_{it}, \nu) = \sum_{j=0}^{\infty} \frac{\lambda_{it}^{j}}{(j!)^{\nu}}\ \&\ \lambda_{it} = \exp(x_{it}^{T}\beta) \quad (1)$$

where the parameter ν corresponds the dispersion index ($\nu < 1$). Using an asymptotic expression for $Z(\lambda_{it}, \nu)$ by Shmueli et al. (2005), where

$$Z(\lambda_{it}, \nu) \simeq \frac{\exp(\nu \lambda_{it}^{\frac{1}{\nu}})}{\lambda_{it}^{\frac{\nu-1}{2\nu}} (2\pi)^{\frac{\nu-1}{2}} \sqrt{\nu}}, \quad (2)$$

$$f(y_{it}) = \frac{\exp[(x_{it}^{T}\beta) y_{it}] \exp[(x_{it}^{T}\beta)(\frac{\nu-1}{2\nu})] (2\pi)^{\frac{\nu-1}{2}} \sqrt{\nu}}{(y_{it}!)^{\nu} \exp(\nu \exp(\frac{x_{it}^{T}\beta}{\nu}))}. \quad (3)$$

$$E(Y_{it}) = \theta_{it} = \lambda_{it}^{1/\nu} - \frac{\nu - 1}{2\nu}\ \ and\ \ Var(Y_{it}) = \frac{\lambda_{it}^{1/\nu}}{\nu} \quad (4)$$

$$Cov(Y_{it}, Y_{i,t-k}) = \rho^{k} (\frac{\theta_{it}}{\nu} + \frac{\nu - 1}{2\nu^{2}}) \quad (5)$$

2.2 Negative-Binomial Longitudinal Model

We assume y_{it} has the negative binomial (NB) distribution with pmf

$$f(y_{it}) = \frac{\Gamma(c^{-1} + y_{it})}{\Gamma(c^{-1}) y_{it}!} (\frac{1}{1 + c\lambda_{it}})^{c^{-1}} (\frac{c\lambda_{it}}{1 + c\lambda_{it}})^{y_{it}} \quad (6)$$

where c is the over-dispersion parameter($c > 0$).

$$E(Y_{it}) = \lambda_{it}, Var(Y_{it}) = \lambda_{it} + c\lambda_{it}^{2}. \quad (7)$$

$$Cov(Y_{it}, Y_{i,t-k}) = \rho^{k} (\lambda_{it} + c\lambda_{it}^{2})(\lambda_{i,t-k} + c\lambda_{i,t-k}^{2}) \quad (8)$$

3 Estimation of Parameters

The Joint Generalized Quasi-likelihood $JGQL$ equation to estimate regression and overdispersion parameters, has a form

$$\sum_{i=1}^{I} D_i^T {\Sigma_i}^{-1}(f_i - \mu_i) = 0, \tag{9}$$

where $f_i = (f_{i1}^T, \ldots, f_{it}^T, \ldots, f_{iT}^T)^T$, $\mu_i = (\mu_{i1}^T, \ldots, \mu_{it}^T, \ldots, \mu_{iT}^T)^T$ are $2T \times 1$ vectors with $f_{it} = (y_{it}, y_{it}^2)^T$, $\mu_{it} = (E(Y_{it}), E(Y_{it}^2))^T$, Σ_i is the covariance matrix of the score vector f_i and D_i is the $2T \times (p+1)$ derivative matrix. The components of this equation are obtained under CP model and NB model to provide $JGQL_{cp}$ and $JGQL_{nb}$ respectively. The $JGQL$ equations are solved iteratively to obtain the estimates of regression and over-dispersion parameters. ρ is consistently estimated using the method of moments.

4 Simulation Study

To compare the performance of the two models discussed in sections 2 and 3, we generate two covariates, such that the first covariate is $x_{it1} = -1(i = 1, \ldots, I/4); 0(i = (I/4) + 1, \ldots, 3I/4); 1(i = (3I/4) + 1, \ldots, I)$ and the second covariate is a set of I standard normal values for all $t = 1, \ldots, T$. We then generate independent clusters of $T = 4$ over-dispersed and correlated counts following AR(1) autocorrelation structure under ***case 1***: Nb model with $\beta_0 = \beta_1 = 1$ and $\rho = 0.9$, for values of c ranging between 0.1 to 3 and ***case 2***: CP model with $\beta_0 = \beta_1 = 0.01$ and $\rho = 0.9$, for values of $\nu = 0.8$ and 0.9. For both the cases, the estimates of β_0, β_1 and c are obtained on the basis of 10,000 simulations under each set, using $JGQL_{nb}$ as well as $JGQL_{cp}$. Also, $\hat{c_{cp}} = \frac{1}{I}\sum_{i=1}^{I} \frac{\lambda_{it}^{\frac{1}{\nu}-2}}{\nu} - \frac{1}{\lambda_{it}}$, where β is the vector of regression parameter estimates under the CP model. The two models are compared on the basis of the standard errors of the estimates (SE) and the actual coverage probability (ACP) and the results are presented in tables 1 and 2.

TABLE 1. JGQL Estimates: Case 1

				NB			CP	
I		$\hat{\beta_{0nb}}$	$\hat{\beta_{1nb}}$	$\hat{c_{nb}}$	$\hat{\beta_{0cp}}$	$\hat{\beta_{1cp}}$	$\hat{\nu_{cp}}$	$\hat{c_{cp}}$
60	$c = 0.1$	0.9827	0.9952	0.1152	1.1011	1.2310	0.9101	0.1230
	SE	(0.2760)	(0.1881)	(0.1885)	(0.2810)	(0.1898)	(0.1901)	(0.1901)
	ACP	0.9665	0.9485	0.9332	0.9121	0.9034	-	0.9021
500		1.1010	0.9999	0.1130	0.8888	1.5333	0.9158	0.1192
	SE	(0.0936)	(0.0944)	(0.0956)	(0.0978)	(0.0985)	(0.0972)	(0.0972)
	ACP	0.9885	0.9787	0.9698	0.9621	0.9623	-	0.9451
60	$c = 3$	0.9818	0.9992	2.9991	1.2101	0.8188	0.2818	3.1011
	SE	(0.1690)	(0.1661)	(0.1774)	(0.1710)	(0.1698)	(0.1790)	(0.1790)
	ACP	0.9497	0.9689	0.9597	0.9189	0.9256	-	0.9222
500		1.1222	0.9885	2.9986	1.3881	0.9212	0.2385	3.0185
	SE	(0.0901)	(0.0918)	(0.0898)	(0.0914)	(0.0956)	(0.0911)	(0.0911)
	ACP	0.9884	0.9892	0.9748	0.9421	0.9543	-	0.9612

TABLE 2. JGQL Estimates: Case 2

		CP			NB			
I		$\hat{\beta_{0cp}}$	$\hat{\beta_{1cp}}$	$\hat{\nu_{cp}}$	$\hat{\beta_{0nb}}$	$\hat{\beta_{1nb}}$	$\hat{c_{nb}}$	$\hat{c_{cp}}$
60	$\nu = 0.8$	0.0110	0.0111	0.7996	0.0109	0.0110	0.1976	0.2014
	SE	(0.1695)	(0.1756)	(0.1305)	(0.1705)	(0.1762)	(0.1320)	(0.1305)
	ACP	0.9697	0.9791	0.9644	0.9612	0.9699	-	-
500		0.0104	0.0103	0.7982	0.0101	0.0103	0.2009	0.1991
	SE	(0.0810)	(0.0756)	(0.0715)	(0.0815)	(0.0760)	(0.0718)	(0.0715)
	ACP	0.9890	0.9902	0.9834	0.9889	0.9799	-	-
60	$\nu = 0.9$	0.0111	0.0112	0.9011	0.0111	0.0114	0.0989	0.1010
	SE	(0.1590)	(0.1325)	(0.1210)	(0.1610)	(0.1362)	(0.1250)	(0.1210)
	ACP	0.9662	0.9745	0.9623	0.9611	0.9699	-	-
500		0.0100	0.0108	0.8985	0.0103	0.0104	0.1024	0.0992
	SE	(0.1001)	(0.0955)	(0.0925)	(0.1011)	(0.0962)	(0.0936)	(0.0925)
	ACP	0.9987	0.9932	0.9901	0.9902	0.9910	-	-

5 Conclusion

When the responses are generated following NB based correlation structure, the estimates obtained by fitting NB model have lower SEs and bigger actual ACPs than the respective SEs and ACPs of the estimates obtained by fitting CP model. The same corresponding conclusions are derived when the responses are generated following CP based correlation structure. These conclusions are reasonable. However, the loss in wrongly using NB model is much lesser than wrongly using CP model. Computationally, under both cases, $JGQL_{cp}$ fails 40 % of the simulations whereas $JGQL_{nb}$ fails only 9 % of the simulations. Moreover, the construction and analysis of CP model is more complicated than that of NB model. Hence, between these two models, NB model is a better choice for analyzing longitudinal over-dispersed count data.

References

Famoye, F. Wulu, J. & Singh K.P. (2004). On the generalized Poisson regression model with an application to accident data. *Journal of Data Science*, **2**, 287-295.

Guikema, S.D. & Goffelt, J.P. (2008). A flexible count data regression model for risk analysis. *Risk analysis*, **28**, 213-223.

Jowaheer, V. (2006). Model misspecification effects in clustered count data analysis. *Statistics & Probability Letters*, **76**, 470-478.

Jowaheer, V. & Sutradhar, B.C. (2002). Analysing longitudinal count data with overdipsersion. *Biometrika*, **89** , 389-399.

Jowaheer, V. & Mamode Khan, N.A. (2009). Estimating Regression effects in Com-Poisson Generalized Linear Model. *International Journal of Computational and Mathematical Science*, **3**, 1339-1351.

Shmueli, G., Minka, T., Borle, J. & Boatwright, P. (2005). A useful distribution for fitting discrete data. *Applied Statistics,Journal of Royal Statistical Society*, **2005**, 486-500.

A three-state semi-Markov model for left-, right-, and interval-censored data

Venediktos Kapetanakis[1], Ardo van den Hout[1], Fiona E Matthews[1]

[1] MRC Biostatistics Unit, Institute of Public Health, Cambridge, U.K. E-mail: venediktos.kapetanakis@mrc-bsu.cam.ac.uk

Abstract: This paper investigates the transition intensities in a three-state semi-Markov model that comprises the states 1:"Healthy", 2:"History of Stroke" and 3:"Death". The Markov assumption is relaxed by adjusting the intensity for the transition from state 2 to state 3 for the time spent in state 2. The exact time of a transition from state 1 to state 2 is unknown due to censoring. The calculation of the likelihood is achieved using a piecewise-constant approach for the transition intensities, integrating out all possible times for the transition from state 1 to state 2.

Keywords: Multi-state model; Semi-Markov; Interval censoring; Piecewise-constant transition intensities.

1 Introduction

Multi-state modelling can be used to analyse longitudinal data when the observed outcome is a categorical variable. Kalbfleisch and Lawless (1985) and others have used this approach in medical applications where the different levels of a progressive disease were seen as the states. A common hypothesis in multi-state modelling is that the data satisfy the first order Markov property. This implies that any previous history of the process can be ignored, which may often be inappropriate. Weiss and Zelen (1965) first proposed a semi-Markov model for clinical trials, assuming that the length of stay in a state followed an arbitrary distribution. Improvements on this framework have been published thereafter.

This work presents a way to fit semi-Markov models following a piecewise-constant approach for the transition intensities (Van den Hout and Matthews, 2009) while handling all types of censoring. This new combination of methods provides great flexibility and extends statistical inference. We investigate the transition intensities in a three-state model with states 1:"Healthy", 2:"History of Stroke" and 3:"Death", and relax the Markov assumption by adjusting the intensity for the transition from state 2 to state 3 for the time spent in state 2. This involves the exact time for the transition from state 1 to state 2. When the data are left or interval censored, this time is

unknown and the estimation of the likelihood for every individual can be found by integrating out all possible times for the transition from state 1 to state 2.
The method is applied to a subset of the Medical Research Council Cognitive Function and Ageing Study (CFAS, `www.cfas.ac.uk`) in the UK. We analyse data from 2321 individuals aged 65 and above. During the period from 1991 to 2003, these individuals had up to 9 interviews where they were asked whether they had had a stroke in the past. Personal data for age (A), gender (G: men versus women), levels of education (E: 10 or more versus less than 10 years of education) and smoking (S: current-smoker versus non- or ex-smoker, at age 60) were also collected.

2 The multi-state model

Semi-Markov models are modelled via regression equations for transition intensities. Let q_{ij} be the intensity for the transition from state i to state j, $(i,j) \in \{(1,2),(1,3),(2,3)\}$. We fit the following model:

$$\log(q_{1s}) = X^T b^{1s} \quad (1)$$

$$\log(q_{23}) = X^T b^{23} + \tau T_2 \quad (2)$$

where $s = 2$ or 3, $X^T = (1, A, G, E, S)$ and T_2 is the time spent in state 2. The exact time for the transition from state 1 to state 2 is unknown due to left-, right- and interval-censoring. Furthermore, it is possible that transitions from state 1 to state 2 may not have been observed before death or right-censoring at the end of follow-up.
We use age (A) as the time scale and the following notation is introduced: A_0 (Age at which all individuals are assumed to be healthy), A_b^i (Age at baseline for the i-th individual), $A_{1_N}^i$ (Age at the last time the i-th individual is observed in state 1), $A_{2_0}^i$ (Age at the first time the i-th individual is observed in state 2), A_N^i (Age at the end of the follow-up for the i-th individual). In our application, we assume that $A_0 = 40$. Figure 1, illustrates all possible transition patterns. In order to fit the model described by (1) and (2), the exact time for the transition from state 1 to state 2 is essential. To solve this problem, we integrate out all possible times for this transition in the calculation of the likelihood for every individual.

2.1 Likelihood contributions

The contribution to the likelihood for individuals with pattern A or D is:

$$L_A = \int_0^{A_{20}-A_{1N}} S(A_{1N} - A_0 + z; q_{12}+q_{13}) q_{12} S(A_N - A_{1N} - z; q_{23}) q_{23}\, dz$$

$$L_D = \int_0^{A_N - A_{1N}} S(A_{1N} - A_0 + z; q_{12}+q_{13}) q_{12} S(A_N - A_{1N} - z; q_{23})\, dz + S(A_N - A_0; q_{12}+q_{23})$$

where $S(t;\lambda)$ is the survival probability of an exponential distribution with parameter λ. For all the other data patterns, the likelihood contributions are calculated in a similar way.

2.2 A piecewise-constant approach

For the likelihood, in every pattern we need to calculate quantities of the form $S\big(A_U - A_L; \lambda(A, T_2)\big)$ for some upper and lower age limits A_U and A_L, respectively; and some parameter $\lambda(A, T_2)$, which changes within the interval $[A_L, A_U]$. The parameter $\lambda(A, T_2)$ is equal to $q_{12} + q_{13}$ or q_{23}, if the current state is state 1 or state 2, respectively. We apply a piecewise-constant approach for the transition intensities and the parameter λ, so that $[A_L, A_U]$ is split in small pieces, within which $\lambda(A, T_2)$ is constant. We specify a resolution, h; and starting from the left limit of $[A_L, A_U]$, we split it to as many pieces of length h as we can. The remaining bit has length $0 \leq l < h$. In every subinterval; $\lambda(A, T_2)$ is evaluated at the left subinterval limit. Figure 2 illustrates the method. The quantities we wish to calculate are given by:

$$S\big(A_U - A_L; \lambda(A, T_2)\big) = \left\{ \prod_{i=0}^{k-1} S\big(h; \lambda(A_L + ih, ih)\big) \right\} S\big(l; \lambda(A_L + kh, kh)\big).$$

References

Kalbfleisch, J.D., and Lawless, J.F. (1985). The analysis of panel data under a Markov assumption. *J Am Stat Assoc*, **80**, 863-871.

Van den Hout, A., and Matthews, F.E. (2009). A piecewise-constant markov model and the effects of study design on the estimation of life expectancies in health and ill health. *Stat Methods Med Res*, **18**, 145-162.

Weiss, G.H., and Zelen, M. (1965). A semi-Markov model for clinical trials. *J. Appl. Prob.*, **2**, 269-86.

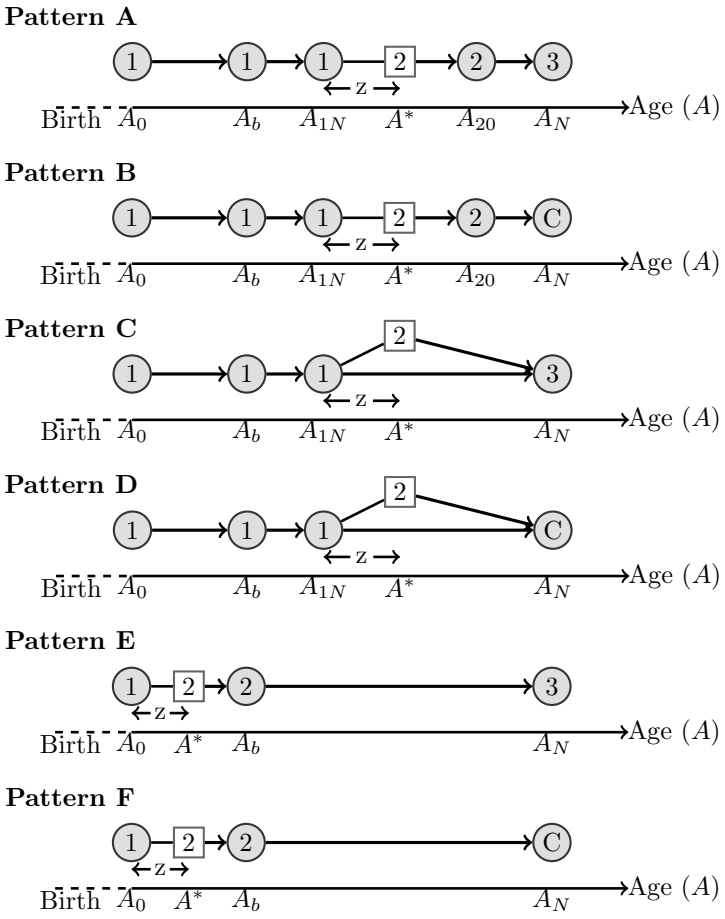


FIGURE 1. Data patterns.

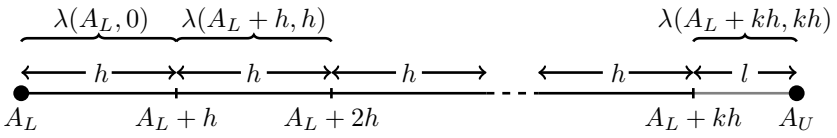


FIGURE 2. Piecewise-constant approach for the calculation of the likelihood.

Assessment of geographical differences in influenza burden using telehealth data: a spatial autoregressive approach

K. Kavanagh[1], C. Robertson[1],[2], J. McMenamin[2]

[1] Department of Mathematics and Statistics, University of Strathclyde, Glasgow, G1 1XH
[2] Health Protection Scotland, Glasgow, G3 7LN

Abstract: A spatial data analysis of the historical counts of calls attributable to influenza-like illness (ILI) received by the National Health Service telephone helpline, NHS24 is presented. Such analysis, broken down by postcode district, assesses the geographical differences in influenza-like illness burden. Spatial association is assessed using Moran's I and spatial variation modelled using Simultaneous Autoregressive (SAR) techniques. This analysis informs whether the use of NHS24 ILI call proportions to assess geographical differences in influenza burden, as has been conducted during the monitoring of pandemic Influenza A vH1N1, is valid.

Keywords: Spatial autoregressive models; NHS24; Influenza.

1 Introduction

During the course of the Influenza A H1N1v pandemic, the Health Protection Agency and its Scottish counterpart, Health Protection Scotland (HPS), have been monitoring levels of influenza in the Scottish community using a suite of surveillance tools. As part of this suite, telehealth data from the Scottish National Health Service (NHS) telephone help line, NHS24 is analysed on a daily basis and published as part of the weekly situation report (HPS, 2009).

NHS24 is a nurse led telephone help line that is the means of access to out of hours general practice services for the 5 million people in Scotland. Each weekday there are about 2500-3000 telephone calls to the service and on weekends and public holidays this rises to about 6000-7000 calls per day. Using the caller's postcode district as a proxy for location, data is routinely analysed at the national and the individual health board level in an exception reporting system framework (Kavanagh *et al.*, 2010).

As part of the weekly situation report, a component of this framework is used to monitor the proportion of all calls made to NHS24 attributable to colds and influenza. Analysis of this data by health board highlights

possible geographical differences in disease burden. Such analysis, however, shows that in certain health boards, for example NHS Lanarkshire, the proportion of total calls attributable to colds and influenza is consistently higher than the Scottish average. This leads to question if there is a systematic reason why ILI call proportions are higher in such health boards, and if so, brings into doubt the unadjusted usage of such a system to compare levels of infectious disease burden between geographical areas.

A possible explanation of geographical differences, not attributable to infection burden, may be differences due to deprivation. It has been observed that deprivation has an effect on call rates to NHS Direct, the English equivalent of NHS24, with higher rates observed in areas where deprivation is at or just above the national average (Burt *et al.*, 2003; Cooper *et al.*, 2005). Such examinations did not however take into account possible spatial autocorrelation in the residual errors which often occurs in population based health data. If such a spatial structure is present it renders classical statistical methods such as general linear modelling, which assume identically distributed errors, invalid.

Within this analysis, we present results of a spatial data analysis of the historical counts of calls attributable to influenza-like illness (ILI) each of the 435 postcode districts in Scotland. This spatial approach accounts for similarities between neighbouring postcode districts and by the use of various socio-demographic and health related variables variables attempts to explain non-infection related geographical differences which may occur in ILI call proportions. Such an analysis will then inform whether the use of NHS24 ILI call proportions to assess geographical differences in influenza burden is valid.

2 Methods

NHS24 call data records information on the age, sex, postcode district and the call reason for each caller to the service. From this data call totals, for each of the 435 postcode districts in Scotland over a year's duration from 1st Sept 2006 - 31st August 2007, are constructed. To assess the levels of ILI in each postcode district a text searching algorithm as described by Kavanagh *et al.* (2010), is used to interrogate the call reason field and classify the call as one of the four syndromes associated with ILI or an alternate other category. A count for the numbers of individuals calling with each ILI related syndrome is then collated for each postcode district. The total count for ILI is an amalgamation of the four syndromes "Colds/Influenza", "Fever", "Difficulty breathing" and "Coughs". The total counts for ILI are then converted into call proportions by dividing by the total number of calls made to NHS24 in the individual postcode district.

Exploratory spatial analysis of the data is achieved through firstly, visualisation of call rates by mapping the data, secondly, formal examination of

the spatial clustering of call rates between regions via Moran's I statistic (Moran, 1950) and construction of a simultaneous autoregressive model to examine the effect area specific covariates on ILI call proportions whilst accounting for spatial autocorrelation between neighbouring postcode districts. The simultaneous autoregressive model takes the following form, outlined by Waller and Gotway (2004),

$$Y(s_i) = \beta x(s_i)' + \sum_{j=1}^{N} b_{ij}\varepsilon(s_j) + \upsilon(s_i) \tag{1}$$

where $x(si)' = (1, x_1(s_i), ..., x_{p-1}(s_i))$ describes the value of each of the $p-1$ covariates in each postcode district $s_i = 1, ..., N$ where $N = 435$. The overall error term is made up of two components, $\sum_{j=i}^{N} b_{ij}\varepsilon(s_j)$ with $b_{ii} = 0$, which describes the dependence between the errors in postcode districts s_i and s_j and $\upsilon(si)$ which is the error in the covariate effects. This formulation then allows us to model the covariate effects which may affect the proportion of calls attributable to ILI in each postcode district whilst controlling for spatial dependence.

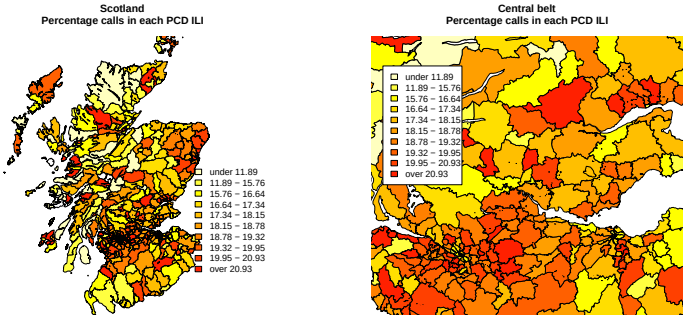
The independent variables considered are; the proportion of individuals in each postcode district aged 0-4 years, the proportion of individuals in each postcode district aged over 65 years, the proportion of individuals in each postcode district with no qualifications, the population density of the postcode district, the deprivation score of the postcode district, the proportion of individuals who rated their health status as "not good" and finally, the proportion of the population with long term limiting illness. All data was taken from the 2001 census (GROS, 2001) and extracted via Casweb (http://casweb.mimas.ac.uk/) with deprivation ranked according to the Carstairs Index (Carstairs and Morris, 1991) and recoded from postcode sector to postcode district by adapting the method of McLoone (2004).

The effect of each covariate is firstly considered individually using a univariate simultaneous autoregressive model eliminating insignificant covariates ($p > 0.3$) prior to building a multivariate model building using backwards elimination. Models are constructed in R using the `spdep` library. As the dependent variable is a proportion, a logit transformation is performed to allow fitting of a linear spatial model via the `spautolm` function.

3 Results

There were 1,366,550 calls made to NHS24 over the year 1st Sept 2006 - 31st August 2007, with 258,805 (18.9%) attributable to ILI. Of these, 17,221 (7%) were classed as Colds/Flu, 52,136 (20%) as Coughs, 91,982 (35%) as difficulty breathing and 97,466 (38%) as Fever.

The proportionately higher burden of these calls in urban areas in Scotland can be seen in Figure 1a with the darker shades indicating a higher



(a) (b)

FIGURE 1. Proportion of all calls which are ILI mapped for (a) every postcode district in Scotland (b) Central belt in detail.

proportion of calls attributable to ILI in the areas surrounding the major cities. This is can also be seen in Figure 1b by very little colour variation within the highest densely populated region in Scotland, the Central Belt. It also appears neighbouring areas, in general, had similar response rates. This spatial similarity is assessed using Moran's I statistic, which indicates a small but statistically significant ($I = 0.09$, p-value=0.005) level of positive spatial autocorrelation present within the data.

Screening of the variables which may have an influence on the proportion of call in each postcode district attributable to influenza-like illness, indicate that the proportion of individuals aged 0-4 years and the proportion over 65 years of age did not have a significant effect ($p = 0.875$ and $p = 0.673$, respectively). All other remaining factors are included in the multivariate model selection summarised in Table 1 which, using the method of backwards selection, results in a final model with only two significant factors; population density and the proportion of individuals working in agriculture. This model illustrates that the proportion of calls attributable to influenza-like illness increases in postcode districts with high population densities and decreases in the more rural areas (using the proportion of individuals working in agriculture as a proxy for rurality).

4 Discussion and conclusions

This study provides evidence that there is spatial clustering in the distribution of the proportion of calls attributable to ILI in the postcode districts in Scotland. This implies that neighbouring postcode districts are more likely to have similar ILI call proportions than those further apart, an intuitive

TABLE 1. Summary of multivariate SAR model selection using backwards elimination.

Original model		
Variable	Estimate (SE)	p-value
Intercept	-3.14 (0.43)	< 0.001
Proportion with "not good" health	5.90(14.73)	0.689
Proportion with no qualifications	3.86 (4.30)	0.369
Carstairs Score	-0.23 (0.12)	0.061
Proportion with limiting long term illness	7.5 (15.43)	0.167
log(Population density)	0.56 (0.17)	0.001
Proportion working in agriculture	-25.30 (6.89)	< 0.001
Final model		
Variable	Estimate (SE)	p-value
Intercept	–2.91 (0.41)	< 0.001
log(Population density)	0.51 (0.16)	0.002
Proportion working in agriculture	-25.92 (6.93)	< 0.001

result given that ILI symptoms generally are attributable to infectious diseases. Therefore if there is a larger infectious disease burden in one area, an neighbouring area is likely to have a higher burden also.

Taking this spatial structure into account, this analysis considered the effect that other area specific covariates may have on ILI burden. Variables which were hypothesised to be potentially relevant were extracted from the 2001 census and their effect examined. The age distribution, the level of education and the self rated health status of individuals in the postcode district were not found to have a significant effect.

The model selected has dependence only on population density and rurality. Including deprivation as a covariate, does not give a statistically significant effect but the proportion of individuals with limiting long term illness has a stronger influence. It is, however, likely that these effects are confounded with population density with the highest proportions of calls attributable to ILI occurring in areas which are most deprived and with a larger proportion of people with long term limiting illness. Half the postcode districts identified as having high ILI call proportions, high population density and low proportion of people employed in agriculture, are in NHS Greater Glasgow and Clyde. Given that 30% of NHS Greater Glasgow and Clyde population are contained within the most deprived 7% of the Scottish population (McLoone, 2004) and deprivation is often linked with ill health, this is in agreement with the finding that this health board has typically high call proportions.

Overall, the model based upon population density and rurality was deemed to explain the variation in the NHS24 ILI call proportions. This implies

that regions are only comparable if they have similar population density and are both rural or both urban. This makes comparison ILI call proportions in different NHS Health Boards difficult as the geographical nature of the health boards implies that they can have very different compositions of rural and urban areas and different overall population densities. This therefore implies that caution must be used when interpreting differences in ILI call proportions between different health boards, as particular health boards may have increased ILI call proportions which are not associated with higher infectious disease burdens. The use of the such systems to compare temporal changes in ILI call proportions within health boards does, however, remain an efficient tool in tracking changes of infection within an area, particularly in the event of increased activity such as the Influenza A H1N1v pandemic.

References

Burt, J. Hooper, R. and Jessopp, L. (2003) The relationship between use of NHS Direct and deprivation in southeast London: an ecological analysis. *Journal of Public Health Medicine*, **25:174-176**

Carstairs, V., and Morris, R. (1991) *Deprivation and Health in Scotland.* Aberdeen University Press.

Cooper, D., Arnold, E., Smith, G., Hollyoak, V., Chinemana, F., *et al.* The effect of deprivation, age and sex on NHS Direct call rates. *British Journal of General Practice*,**55: 287-291**

GROS (2001) `http://www.gro-scotland.gov.uk/census/censushm/index.html`, accessed online 4 Feb 2009

HPS (2009). Weekly Situation Report Influenza A H1N1v. `http://www.hps.scot.nhs.uk/resp/swineinfluenzareports.aspx`, accessed online 4 Feb 2009

Kavanagh, K., Murdoch, H., Robertson, C. and McMenamin, J. (2010) Development of an exception reporting system for 'influenza-like' syndromes in Scotland and usage during the Influenza A H1N1v pandemic. *In preparation.*

McLoone P. (2004) Carstairs scores for Scottish postcode sectors from the 2001 Census. `http://www.sphsu.mrc.ac.uk/files/File/library/other%20reports/Carstairs_report.pdf`, accessed online 4 Feb 2009

Moran, P.A.P. (1950) Notes on Continuous Stochastic Phenomena. *Biometrika*, **37: 17-33**

Waller, L.A. and Gotway, C.A. (2004) *Applied Spatial Statistics for Public Health Data.* John Wiley & Sons.

Spatial clustering of TB-infected cattle herds in Ireland prior to and following proactive badger removal

Gabrielle E. Kelly[1]

[1] School of Mathematical Sciences, University College Dublin, Belfield, Dublin 4, Ireland.

Abstract: Data from the Four Area Project, a large-scale intervention study aimed at assessing the effect of proactive badger culling on bovine TB incidence in cattle herds, are analyzed using logistic models that explicitly include spatial random effects i.e. generalized linear geostatistical models (Diggle and Ribeiro (2007)). The models indicate that infected herds are spatially correlated and the scale at which spatial correlation occurs is established and is shown to vary with time. It is also found that this spatial correlation may persist in areas where badgers have been removed. The results are useful in informing TB control policy.

Keywords: geostatistical models; bovine TB; badger; practical range.

1 Introduction

Bovine TB (*Mycobacterium bovis*) is a disease that affects cattle and badgers (*Meles meles*). The annual herd incidence is around 6% in Ireland (http://www.agriculture.gov.ie) while incidence in the UK is roughly double this (http://www.defra.gov.uk). Two large-scale field trials, the Four Area Project (FAP) in Ireland and the Randomized Badger Culling Trial (RBCT) in England, presented strong evidence that badgers infect cattle (Griffin et al. (2005), Bourne et al. (2007)). Both trials involved comparisons between large areas where badgers were pro-actively culled to reference areas where little or no culling was carried out. The RBCT also included comparisons between areas with no culling and areas where reactive culling (in response to TB outbreaks) was carried out and concluded that such culling leads to an increase in bovine TB due to a perturbation of badger habitats. In Ireland reactive culling typically occurs in the index and neighboring farms - a distance of a few kilometres, while proactive culling in the FAP and RBCT was in areas of well over 100 km^2. The studies have contributed to an intense national debate in England regarding the role of badger culling in the control of the disease in cattle. Emerging from the debate is the question of the distances at which badger-to-cattle transmission occurs and the scale at which badger culling is likely to be effective. This

TABLE 1. Number of herds and percentage with at least two cattle TB positive in the reference and removal areas in four counties in Ireland during September 1992 to August 1997 (period 0) and September 1997 to August 2002 (period 1).

Area	Cork rem	Cork rem	Cork ref	Cork ref	Don rem	Don rem	Don ref	Don ref
Period	0	1	0	1	0	1	0	1
No. tested	300	292	278	282	404	391	390	370
% positive	27	15	21	28	7	0	3	6
Area	Kilk rem	Kilk rem	Kilk ref	Kilk ref	Mon rem	Mon rem	Mon ref	Mon ref
Period	0	1	0	1	0	1	0	1
No. tested	244	240	242	234	712	720	568	583
% positive	17	10	13	25	11	6	16	18

is also of import regarding the current development of vaccines for badgers and cattle. Here, using data gathered in the five year periods, before (1992-1997) and during (1997-2002) proactive culling in lands of the FAP, we establish practical spatial correlation ranges at which TB clustering takes place in cattle and the effect if any badger removal had on these.

2 Materials and Methods

2.1 Cattle and badger population structures

Herds that had all their land contained in the removal or reference areas of the project and were tested at least annually, as in Griffin et al. (2005), are included in the analysis. The geographical location of a herd is taken as the centroid of the largest land fragment owned by the farmer and recorded in a GIS database. A herd is designated infected in a year if it contained at least two cattle that tested positive that year. The data are summarized in Table 1. In the removal areas period 1, large numbers of badgers were removed particularly in the first year. We also include in the analysis the 1,039 adult badgers culled in that first year for which the infection status is known. Overall 20% of badgers were infected with TB, ranging from 13% in Kilkenny to 41% in Donegal.

2.2 Spatial models

We consider marginal logistic regression spatial models for the binary responses $Y(s_i)$ (ith herd at location s_i infected/not-infected), with mean

$$E(Y(s)) = \mu = g^{-1}(X\beta),$$

where $g(\mu) = \text{logit}(\mu)$ and X is the matrix of covariates. Let $u(s_i)$ be a spatial random effect at location s_i. We assume the $u(s_i)$ follow an exponential

isotropic covariance model F with (i,j)th element given by

$$\mathrm{Cov}[u(s_i), u(s_j)] = \sigma^2[\exp(-d_{ij}/\rho)]$$

where d_{ij} is the distance between the locations s_i and s_j. The variance-covariance matrix of the data is modeled as

$$\mathrm{Var}[Y(s)] = A^{1/2}FA^{1/2}$$

where A is a diagonal matrix with elements $\mu_i(1-\mu_i)$. The parameters σ^2 and ρ refer to the geostatistical parameters sill and "range", respectively. Covariance in this model reaches zero only asymptotically, thus the practical range is defined as the distance at which covariances are reduced to 5% of the sill i.e. 3ρ. A nugget effect is included by using

$$\mathrm{Var}[Y(s)] = c_0A + A^{1/2}FA^{1/2}$$

as in Schabenberger and Gotway (2005). Estimation in a spatial model involves repeated inversion of an $n \times n$ matrix, where n is the number of distinct geographical locations, which is computationally difficult to implement. Also, the scale at which spatial correlation occurs in each treatment area and county varies considerably. Moreover, as infection rates are relatively low, sparse data problems arise when data is considered yearly. Therefore, separate spatial models are fitted by county, area and period.
The fixed effects in the models were *log(herd size)*, *ph* (presence or absence of previous infection in the herd) and the factor *year* representing the years 1992-2002. In the removal areas period 1, the covariates, *d* (distance to the nearest badger sett), *infb* (infection status of nearest badger sett) and two and three-way interactions of *ph*, *infb* and *d* were also considered.
Models are fitted using the GLIMMIX procedure in the software package SAS version 9.1.3 (SAS Institute Inc., Cary, North Carolina, USA)

3 Results

In the sixteen fitted models the covariate *log(herd size)* was statistically significant ($p < 0.001$) in all models. Differences between years were found in some models according as whether or not there were large changes in rates of infection over time as shown in Table 1. In the removal areas period 1 models (the number infected in Donegal was too low to support modeling), in Cork and Monaghan for herds with the nearest badger sett infected, there was no difference between those with a previous history and those without; and for herds with the nearest badger sett not infected, previous history was a risk factor for disease (p=0.003 Cork and p=0.002 Monaghan). *ph* had no significant effect in any other models.
For periods 0 and 1 respectively, in the Cork removal area the practical range is 4.40 km and 4.76 km; in Monaghan removal it is 0 km and 2.90

km; it is 0 km and 0 km in Kilkenny; and 9.13 km in Donegal for period 0. For periods 0 and 1, in the Cork reference area the practical range is 2.70 km and 3.64 km; in Monaghan reference it is 5.15 km and 10.04 km; in Kilkenny reference it is 2.48 km and 8.85 km; and in Donegal reference 0.00 km and 5.52 km. In cases where there are long distance differences across the area quite a high value of the range parameter ρ will fit and be very unstable. It indicates broad spatial heterogeneity. An estimate of $\rho = 0$ also fits well in such cases - as in Monaghan removal period 0. In all models there was no nugget effect.

4 Conclusions

Spatial clustering of infected herds is found for almost all periods and areas within a county. Importantly, we note there is variation in range between counties. This is useful information for culling policy. The variation may be associated with varying rates of bovine TB over time, the mode of transmission, the farming environment and badger ecology.
It is noteworthy that spatial association of infection persists during the proactive badger culling period in Cork and Monaghan. This and the result in Cork and Monaghan with respect to previous history, may indicate sources of transmission of disease other than the badger.

References

Bourne, J., Donnelly, C., Cox, D., Gettinby, G., McInerney, J., Morrison, I. and Woodroffe, R. (2007). Bovine TB: the Scientific Evidence. A Science Base for a Sustainable Policy to Control TB in Cattle. Final Report of the Independent Scientific Group on Cattle TB. Department of Environment Food and Rural Affairs, London, UK. http://www.defra.gov.uk/foodfarm/farmanimal/diseases/atoz/tb/isg/report/final report.pdf

Diggle, P.J., and Ribeiro, P.J. (2007). *Model-based Geostatistics*. New York: Springer.

Griffin, J.M., Williams, D.H., Kelly, G.E., Clegg, T.A., O'Boyle, I., Collins, J.D. and More, S.J. (2005). The impact of badger removal on the control of tuberculosis in cattle herds in Ireland. *Preventive Veterinary Medicine*, **67**, 237-266.

Schabenberger, O. and Gotway, C.A. (2005). *Statistical methods for spatial data analysis*. London: Chapman & Hall/CRC.

Cluster analysis for joint continuous and discrete correlated data

Arnošt Komárek[1]

[1] Dept. of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University in Prague, Sokolovská 83, 186 75 Praha 8, Czech Republic. E-mail: Arnost.Komarek@mff.cuni.cz

Abstract: We introduce a multivariate generalized linear mixed model with a normal mixture in the random effects distribution and show how to use the proposed model for a cluster analysis based on grouped continuous and discrete data obtained jointly on subjects being clustered. For practical analysis we present an extension of the R package `mixAK`.

Keywords: Generalized linear mixed model; Longitudinal data; Markov chain Monte Carlo; Normal mixture; R package.

1 Introduction

In classical cluster analysis, the subjects are grouped into clusters on basis of their characteristics represented by i.i.d. random variables/vectors $\mathbf{Y}_i$ $(i = 1, \ldots, N)$ where N stands for the number of subjects. Nevertheless, in many practical situations, the outcomes we want to use for clustering do not fit in such an i.i.d. framework. Two such situations, frequently encountered in practice include: (a) longitudinal studies in which each subject is represented by a distinct number of measurements performed at irregular time points being different between subjects (see Section 1.1), (b) studies in which different numbers of repeated measurements (not necessarily in time) are recorded for each subject leading to hierarchically structured data (see Section 1.2).

Moreover, in longitudinal studies, different outcomes (both discrete and continuous) are often recorded at each visit and longitudinal evolution of all available outcomes can be used for clustering. Similarly, in the case of hierarchical data, each subject can be represented by different outcomes recorded on subunits and we aim to use all available measurements for clustering. In the rest of the paper, we present a clustering method based on a multivariate generalized linear mixed model (GLMM) fitted to the measured outcomes. The clustering procedure will be based on assumed finite mixture of normal distributions for the random effects in the GLMM. To make the estimation procedure computationally feasible, we exploit a Bayesian specification of the model and base the inference on the Markov

chain Monte Carlo (MCMC) simulation. To clarify the matters, we consider the following example data.

1.1 The Czech growth data

In a period 1997–2000, a group of 1 944 (1 005 boys and 939 girls) Czech children aged 6–16 was longitudinally followed (at most 6 visits designed to happen every 6th month) in a study conducted to explore somatic development of contemporary children (Bláha et al., 2006). Among other things, weight and height was recorded at each visit. Further, for each child, his/her body mass index (BMI) was computed at each visit and compared to the 90-th population gender and age specific percentile (obtained from other cross-sectional studies). The BMI value exceeding this percentile was considered as indication of obesity problems. For several practical reasons, it is of interest to identify groups of typical development of somatic measures (like weight and height) together with development of exhibition of obesity among children. In our first illustration, we show the cluster analysis based on longitudinal (and hence correlated) observations of two continuous (weight and height) and one binary (obesity – yes/no) outcomes.

1.2 The NTP data

Price et al. (1995) report a study performed in the framework of the National Toxicity Program (NTP) of the National Institute of Health in which timed-pregnant CD-1 mice were dosed by gavage with Ethylene Glycol (EG, 0, 750, 1 500, 3 000 mg/kg/day) in distilled water. Dosing occurred during the period of organogenesis and structural development of the foetuses (gestational days 6 through 15). Following sacrifice at gestational day 17, the uterine contents of each of 94 dams were evaluated for the number of live fetuses (1 – 16 per dam), among other things. Further, each live fetus was examined for evidence of malformations and its fetal weight was recorded. For our illustration, each dam is represented by a vector of 1 – 16 continuous (fetal weight) and 1 – 16 binary (fetus malformed – yes/no) outcomes. Furthermore, all outcomes taken on a single dam should be considered to be correlated. In our second illustration, we show how to cluster dams on basis of these grouped continuous and binary outcomes.

2 Model

2.1 Multivariate generalized linear mixed model with a normal mixture in the random effects distribution

Let us first introduce some notation. For the ith subject ($i = 1, \ldots, N$), let $\mathbf{Y}_{i,r} = (Y_{i,r,1}, \ldots, Y_{i,r,n_{i,r}})'$ denote a random vector of the rth response ($r = 1, \ldots, R$). For example, for the Czech growth data, $R = 3$,

$n_{i,1} = n_{i,2} = n_{i,3} \in \{1, \ldots, 6\}$ with $\mathbf{Y}_{i,1}$, $\mathbf{Y}_{i,2}$, $\mathbf{Y}_{i,3}$ representing longitudinally measured weights, heights and indicators of obesity, respectively of the ith child. For the NTP data, $R = 2$, $n_{i,1} = n_{i,2} \in \{1, \ldots, 16\}$ with $\mathbf{Y}_{i,1}$ and $\mathbf{Y}_{i,2}$ representing fetal weights and malformation statuses, respectively of the fetuses of the ith dam. Further, let $\mathbf{Y}_i = (\mathbf{Y}'_{i,1}, \ldots, \mathbf{Y}'_{i,R})'$ be a random vector of all measurements on the ith subject. Finally, let $\mathbf{Y} = (\mathbf{Y}'_1, \ldots, \mathbf{Y}'_N)'$ be a random vector of all available outcomes.
First, we model each response using a standard GLMM

$$\begin{aligned} h_r^{-1}\big\{\mathsf{E}(Y_{i,r,j} \,|\, \alpha_r,\, \mathbf{b}_{i,r})\big\} \;&=\; \mathbf{x}'_{i,r,j}\alpha_r + \mathbf{z}'_{i,r,j}\mathbf{b}_{i,r}, \\ &\qquad i = 1, \ldots, N,\; r = 1, \ldots, R,\; j = 1, \ldots, n_{i,r}, \end{aligned} \tag{1}$$

where h_r^{-1} is the link function for the rth response, $\mathbf{x}_{i,r,j}$, $\mathbf{z}_{i,r,j}$ are vectors of known covariates which may include a constant for intercept, time values for longitudinal observations or any other additional covariates (e.g., gender), α_r is a vector of unknown regression coefficients. Finally, $\mathbf{b}_{i,r}$ is a vector of random effects for the rth response specific for the ith subject. Let $\mathbf{b}_i = (\mathbf{b}'_{i,1}, \ldots, \mathbf{b}'_{i,R})'$ be a joint vector of random effects from all R responses for the ith subject.
Being within the GLMM framework, we assume that $p(Y_{i,r,j} \,|\, \alpha_r,\, \mathbf{b}_{i,r})$, the conditional distribution of $Y_{i,r,j}$ given α_r and $\mathbf{b}_{i,r}$ belongs to an exponential family with the mean specified by (1). Further, we assume that given $\mathbf{b}_i$, the R response vectors $\mathbf{Y}_{i,1}, \ldots, \mathbf{Y}_{i,R}$ are independent for each $i = 1, \ldots, N$. Finally, we assume that random vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ representing different subjects are independent. Dependence between the R response vectors $\mathbf{Y}_{i,1}, \ldots, \mathbf{Y}_{i,R}$ is modelled by a joint distribution for the random effect vector which we specify as a normal mixture, i.e., we assume that the vectors $\mathbf{b}_1, \ldots, \mathbf{b}_N$ are i.i.d. with a density

$$p(\mathbf{b} \,|\, \theta) \;=\; \sum_{k=1}^{K} w_k\, \varphi(\mathbf{b} \,|\, \mu_k,\, \mathbf{D}_k), \tag{2}$$

where K is pre-specified number of mixture components, $\varphi(\cdot \,|\, \mu_k, \mathbf{D}_k)$ is a density of the (multivariate) normal distribution with mean μ_k and a covariance matrix $\mathbf{D}_k$ and $\theta = (w_1, \ldots, w_K, \mu'_1, \ldots, \mu'_K, \mathsf{vec}(\mathbf{D}_1), \ldots, \mathsf{vec}(\mathbf{D}_K))'$ is a vector of unknown parameters. In subsequent clustering, each mixture component will represent one cluster. Further, it is useful to introduce for each subject latent random variables $U_1, \ldots, U_N \in \{1, \ldots, K\}$. The mixture model (2) can then be written hierarchically as

$$\left.\begin{aligned} p(\mathbf{b} \,|\, \theta,\, U = k) \;&=\; \varphi(\mathbf{b} \,|\, \mu_k,\, \mathbf{D}_k) \\ \mathsf{P}(U = k \,|\, \theta) \;&=\; w_k \end{aligned}\right\} k = 1, \ldots, K. \tag{3}$$

With $R = 1$, the proposed model was introduced as the heterogeneity model by Verbeke and Lesaffre (1996, continuous response only), and by

Molenberghs and Verbeke (2005, Sec. 23.2, response following the exponential family distribution). They considered maximum-likelihood estimation through the EM algorithm and in their practical illustrations, the authors limited themselves to the homoscedastic normal mixture ($\mathbf{D}_1 = \cdots = \mathbf{D}_K$) and bivariate random effects corresponding to random intercept and slope for modelled response. To make the estimation computationally feasible also for more complex models (typically resulting from situations with $R > 1$ and responses of different nature), we opted for a Bayesian specification of the model and inference based on the MCMC simulation.
That is, having specified the prior distribution $p(\theta)$ for the model parameters, we obtain a random sample $\theta^{(m)},\ u_1^{(m)}, \ldots, u_N^{(m)},\ \mathbf{b}_1^{(m)}, \ \ldots, \ \mathbf{b}_N^{(m)}$, $m = 1, \ldots, M$ from the posterior distribution $p(\theta, u_1, \ldots, u_N, \mathbf{b}_1, \ \ldots, \mathbf{b}_N \,|\, \mathbf{y})$ of the model parameters and latent quantities (component allocations and model parameters) and base the inference on this random sample. The prior distribution is specified in a weakly informative way in an analogous manner to that described in Komárek et al. (2010) in the context of the linear mixed model. The MCMC sampling is implemented using the block Gibbs algorithm with Metropolis-Hastings steps in situations in which the full conditional distribution does not have a standard form. Finally, to be able to use the MCMC output for subsequent clustering, we resolved a known label switching problem using algorithms given by Stephens (2000). For practical computation, the R package `mixAK` (Komárek, 2009) has been extended and can be downloaded from CRAN.

2.2 Clustering procedure

With the hierarchical specification (3) of the model and within the Bayesian framework, the clustering is naturally based on posterior probabilities $\pi_{i,k} = \mathsf{P}(U_i = k \,|\, \mathbf{y})$, $i = 1, \ldots, N$, $k = 1, \ldots, K$ which for each subject express the strength of belonging to each cluster. If a terminate classification is of interest, we may classify the ith subject in the gth group if and only if $\pi_{i,g} = \max_{k=1,\ldots,K} \pi_{i,k}$ or if and only if $\pi_{i,g} = \max_{k=1,\ldots,K} \pi_{i,k}$ exceeds a certain threshold probability δ ($0 < \delta < 1$) in which case a certain proportion of subjects may remain unclassified.
Exploiting the fact that

$$\mathsf{P}(U_i = k \,|\, \mathbf{y}) = \mathsf{E}_{U_i}\big\{\delta(U_i = k) \,\big|\, \mathbf{y}\big\} \tag{4}$$

or

$$\mathsf{P}(U_i = k \,|\, \mathbf{y}) = \mathsf{E}_{\theta,\mathbf{b}_i}\big\{\mathsf{P}(U_i = k \,|\, \theta, \mathbf{b}_i, \mathbf{y}) \,\big|\, \mathbf{y}\big\} \tag{5}$$

it is relatively easy to obtain MCMC based estimates for $\pi_{i,k}$, $i = 1, \ldots, N$, $k = 1, \ldots, K$.

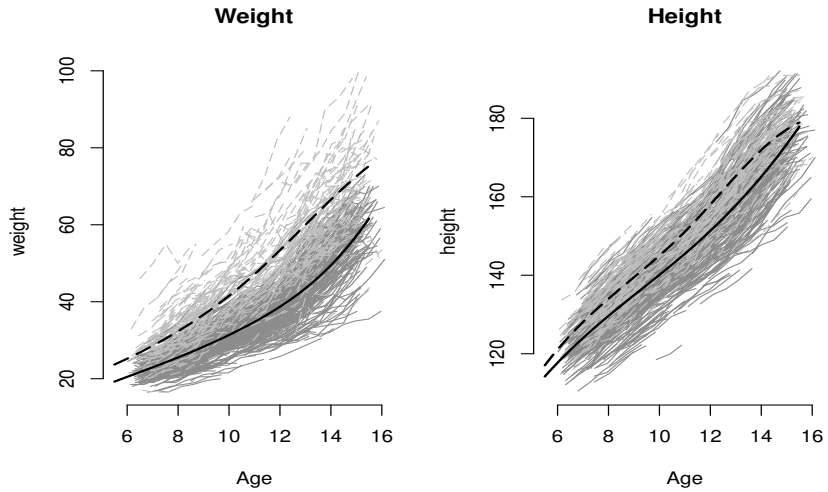


FIGURE 1. Observed and fitted (based on posterior means) profiles of weight and height in the two clusters. Black lines: fitted profiles, grey lines (darker – cluster with $\hat{w}_1 = 0.73$, brighter – cluster with $\hat{w}_2 = 0.27$): observed profiles.

3 Illustration: The Czech growth data

As illustration, we analyzed boys from the Czech growth data introduced in subsection 1.1. The model (1) has been fitted to continuous responses weight and height and dichotomized BMI (one if it exceeds the population 90-th percentile). For both weight and height a B-spline of order 3 with knots at 5.5, 11.0, and 16.5 has been used to describe the evolution of weight and height over time. All (five) B-spline coefficients have been assumed to be random for both weight and height. For dichotomized BMI, a simpler random intercept and slope model has been assumed. In total, there are 12-dimensional (5+5+2) random effects in the model which serve as a basis for clustering in which we assumed $K = 2$ components.

The posterior means of the mixture weights (estimated cluster probabilities) are 0.73 and 0.27. Using the simple classification rule which allocates the i-th boy to cluster k if $\pi_{i,k} = \mathsf{P}(U_i = k \,|\, \mathbf{y}) > 0.5$, 730 boys were classified in group 1 and 273 in group 2. Figure 1 shows observed profiles (different hues of grey for boys classified in the two clusters) and fitted profiles of continuous responses based on posterior means of model parameters. The smaller cluster ($\hat{w}_1 = 0.27$) includes heavier and taller boys and also all boys who got over the BMI 90-th percentile at least once. That is, one could consider this cluster as being representative for boys who stay at higher risk for obesis. Of course, it is also possible to characterize the

smaller cluster as a group of faster growing boys. Nevertheless, the hypothesis that the smaller cluster is at least partially representative for boys being at higher risk for obesis is also supported by the fact that the weight difference (based on fitted profiles) between the two groups increases over the whole period, nevertheless the height difference reaches its maximum around the age of 13 (peak of pubescence) and then starts to decrease.

Acknowledgments: The work on this paper has been supported by the grant GAČR 201/09/P077, Czech Science Foundation and the grant MSM 0021620839, Ministry of Education, Youth and Sports of the Czech Republic.

References

Bláha, P., Krejčovský, L., Jiroutová, L., Kobzová, J., Sedlák, P., Brabec, M., Riedlová, J., and Vignerová, J. (2006). *Somatický vývoj současných českých dětí. Semilongitudinální studie (Somatic Development of Contemporary Czech Children. Semilongitudinal Study).* Praha: Univerzita Karlova v Praze, katedra antropologie a genetiky člověka, Státní zdravotní ústav Praha.

Komárek, A. (2009). A new R package for Bayesian estimation of multivariate normal mixtures allowing for selection of the number of components and interval-censored data. *Computational Statistics and Data Analysis*, **53**, 3932–3947.

Komárek, A., Hansen, B. E., Kuiper, E. M. M., van Buuren, H. R., and Lesaffre, E. (2010). Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution. *Statistics in Medicine*, **29**. To appear, doi: 10.1002/sim.3849.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* New York: Springer.

Price, C.J., Kimmel, C.A., Tyl, R.W., and Marr, M.C. (1985). The developmental toxicity of ethylene glycol in rats and mice. *Toxicology and Applied Pharmacology,* **81**, 113–127.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, **91**, 217–221.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B*, **62**, 795–809.

Identifying subtypes of Acute Myeloid Leukemia: A model based approach

Matthias Kormaksson[1]*, Maria Figueroa[2], James Booth[1]

[1] Department of Statistical Science, Cornell University, Ithaca, NY 14853; *Communicating Author: <mk375@cornell.edu>.

[2] Department of Medicine, Hematology Oncology Division, Weill Cornell Medical College, 525 East 68th Street, New York, NY 10065.

Abstract: A recent study on a cohort of 344 well-characterized patients with acute myeloid leukemia suggests that subjects can be segregated into distinct groups using unsupervised clustering based on their DNA methylation profiles. We suggest a model based approach, where we introduce latent cluster specific methylation indicators on each gene. These indicators along with some standard assumptions impose a specific mixture distribution on each cluster and the parameters of the induced model are estimated using the EM algorithm. The model can be extended by introducing latent gene importance indicators, which provides us with a powerful tool for classification.

Keywords: Acute myeloid leukemia (AML); methylation; clustering; classification; EM-algorithm.

1 Introduction

The data, which involves methylation profiles of $n = 344$ patients with AML, was collected at Erasmus University Medical Center (Rotterdam) between 1990-2008, see Figueroa et al. (2010) The main goal of Figueroa et al. (2010) was to cluster these patients in order to identify distinct subtypes of AML. They used simple hierarchical clustering methods using correlation similarity, which gave some promising results. We suggest constructing a likelihood for any given partition of the patient set $\{1, \ldots, n\}$ and use it as a base for clustering and classification of patients.

1.1 Clustering

Let y_{ij} denote the methylation response for patient $i = 1, \ldots, n$, and gene $j = 1, \ldots, G$, where $G = 25,626$. Looking at a histogram of $(y_{ij})_{j=1,\ldots,G}$ for each patient i we see roughly a mixture of two normals, where the left mode corresponds to methylated genes and the right mode to genes that are not methylated. As many of the genes behave almost identically across subjects we only focus on a subset of genes $J_d \subset \{1, \ldots, G\}$, in an attempt

to eliminate noise. Figueroa et al. (2010) let J_d be the set of all genes with standard deviations > 1. Let $\mathcal{C}$ be a partition of the n subjects. We assume patients in a given cluster $c \in \mathcal{C}$ have identical methylation profiles and introduce latent indicators $(w_{cj})_{j \in J_d}$ that are defined as follows:

$$w_{cj} = \begin{cases} 1 & \text{if gene } j \text{ is methylated in cluster } c \\ 0 & \text{if gene } j \text{ is not methylated in cluster } c \end{cases} \tag{1}$$

We assume a priori that the vector of these latent indicators, $\mathbf{w}$, has density

$$f(\mathbf{w}) = \prod_{c \in \mathcal{C}} \prod_{j \in J_d} \pi_{1c}^{w_{cj}} \pi_{2c}^{1-w_{cj}}, \quad \pi_{1c} + \pi_{2c} = 1$$

and the conditional distribution of the observed data, $\mathbf{y}$, given $\mathbf{w}$ is

$$f(\mathbf{y}|\mathbf{w}) = \prod_{c \in \mathcal{C}} \prod_{j \in J_d} \Big(\prod_{i \in c} \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2) \Big)^{w_{cj}} \Big(\prod_{i \in c} \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2) \Big)^{1-w_{cj}}, \tag{2}$$

where ϕ denotes the normal density. Note that histograms for different patients are not necessarily on the same scale (array effect) and hence we allow for individual specific parameters. We have the restriction $\mu_{1i} < \mu_{2i}$ for all i, since lower values of the response indicate methylation. This construction leads to the mixture likelihood

$$L_{\mathcal{C}} = \prod_{c \in \mathcal{C}} \prod_{j \in J_d} \left(\pi_{1c} \prod_{i \in c} \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2) + \pi_{2c} \prod_{i \in c} \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2) \right). \tag{3}$$

The above likelihood is maximized using the EM-algorithm.
The goal is now to find the partition that leads to the highest maximized likelihood, $L_{\mathcal{C}}$. Since calculating $L_{\mathcal{C}}$ for all partitions $\mathcal{C}$ is computationally impossible, we suggest a simple hierarchical algorithm that starts by maximizing (3) when each patient represents his/her own cluster. Then we merge the two patients/clusters that leads to the highest value of (3). We continue merging clusters under this maximum likelihood criteria until we are left with one big cluster. We finally pick the partition that has the highest value of $L_{\mathcal{C}}$. Note that including an AIC penalty makes little difference as we are only reducing the number of parameters by one each time we merge two clusters. Also note that this method determines the number of clusters automatically.

1.2 Classification

In the classification setting we have data on n patients and we know which class each patient belongs to. Given new subjects, we wish to come up with a discriminant rule to determine their AML classification. Let us assume there is a subset $J_d(\mathcal{C}) \subset \{1, \ldots, G\}$ of genes that discriminate well between

patients. For genes in the complementary set, $\overline{J_d(\mathcal{C})}$, we assume they either methylate for all of the patients or they do not methylate for all patients. We let p denote the proportion of genes in $J_d(\mathcal{C})$ and assume the observed data, $\mathbf{y}$, follows the mixture density

$$\begin{aligned} f(\mathbf{y}) &= \prod_{j=1}^{G} \Bigg\{ p \prod_{c\in\mathcal{C}} \Big(\pi_{1c} \prod_{i\in c} \phi(y_{ij}|\mu_{1i},\sigma_{1i}^2) + \pi_{2c} \prod_{i\in c} \phi(y_{ij}|\mu_{2i},\sigma_{2i}^2) \Big) \\ &\quad + (1-p)\Big(\pi_1 \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{1i},\varsigma_{1i}^2) + \pi_2 \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{2i},\varsigma_{2i}^2) \Big) \Bigg\}. \quad (4) \end{aligned}$$

Note that this is a mixture of two mixture densities, where the first mixture component corresponds directly to the mixture likelihood, $L_{\mathcal{C}}$, in (3) and the second mixture component corresponds to (3) when $\mathcal{C}$ is just one big cluster. We also allow for individual specific parameters for the second mixture component and put the restriction $\alpha_{1i} < \alpha_{2i}$ for all i. In order to maximize (4) using the EM-algorithm, we define w_{cj} as we did in (1), and introduce a latent gene importance indicator, γ_j, which equals 1 if $j \in J_d(\mathcal{C})$ and 0 otherwise. This idea of gene importance indicators was introduced in Tadesse et al. (2005). We put the following priors on $\boldsymbol{\gamma}$

$$f(\boldsymbol{\gamma}) = \prod_{j=1}^{G} p^{\gamma_j}(1-p)^{1-\gamma_j},$$

and the following conditional prior on $\mathbf{w}$

$$f(\mathbf{w}|\boldsymbol{\gamma}) = \prod_{j=1}^{G} \Big(\prod_{c\in\mathcal{C}} \pi_{1c}^{w_{cj}} \pi_{2c}^{1-w_{cj}} \Big)^{\gamma_j} \Big(I(\mathbf{w}_j \in A) \pi_1^{w_{1j}} \pi_2^{1-w_{1j}} \Big)^{1-\gamma_j},$$

where $A = \{\mathbf{w}_j = (w_{cj})_{c\in\mathcal{C}} | w_{cj} = w_{c'j}, \text{ all } c, c' \in \mathcal{C}\}$.
With the EM-algorithm we maximize (4) and also obtain posterior expectations of the latent indicators $(\gamma_j)_j$, and $(w_{cj})_{c,j}$. Classification of a new patient i involves treating these posterior expectations as the true values and focus only on genes j such that $\gamma_j \approx 1$, or $j \in J_d(\mathcal{C})$. The likelihood of a new observation $(y_{ij})_{j\in J_d(\mathcal{C})}$, on the assumption that $i \in c$, is given by (see (2)):

$$L_c = \prod_{j\in J_d} \phi(y_{ij}|\mu_{1i},\sigma_{1i}^2)^{\hat{w}_{cj}} \phi(y_{ij}|\mu_{2i},\sigma_{2i}^2)^{1-\hat{w}_{cj}}.$$

This likelihood can now be maximized and we arrive at a classification rule that assigns patient $i \in c$ if $L_c > L_{c'}$ for all $c' \neq c$.

2 EM Algorithm

In this section we will briefly describe the classification EM algorithm. This algorithm includes the clustering EM algorithm as a special case. The

complete data loglikelihood that leads to the marginal for $\mathbf{y}$ given in (4) is

$$\begin{aligned}
& \log f(\mathbf{y}, \mathbf{w}, \boldsymbol{\gamma}|\boldsymbol{\theta}) \\
= & \sum_{j=1}^{G} \Big\{ \gamma_j \log p + (1-\gamma_j)\log(1-p) \Big\} \\
+ & \sum_{j=1}^{G} \sum_{c\in\mathcal{C}} \Big\{ \gamma_j w_{cj} \log \pi_{1c} + \gamma_j (1-w_{cj}) \log \pi_{2c} \Big\} \\
+ & \sum_{j=1}^{G} \sum_{c\in\mathcal{C}} \Big\{ \gamma_j w_{cj} \sum_{i\in c} \log \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2) + \gamma_j (1-w_{cj}) \sum_{i\in c} \log \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2) \Big\} \\
+ & \sum_{j=1}^{G} (1-\gamma_j) \log \left(I(w_{cj} = w_{c'j}, \text{ all } c, c' \in \mathcal{C}) \right) \\
+ & \sum_{j=1}^{G} \Big\{ (1-\gamma_j) w_{1j} \log \pi_1 + (1-\gamma_j)(1-w_{1j}) \log \pi_2 \Big\} \\
+ & \sum_{j=1}^{G} \Big\{ (1-\gamma_j) w_{1j} \sum_{i=1}^{n} \log \phi(y_{ij}|\alpha_{1i}, \varsigma_{1i}^2) + (1-\gamma_j)(1-w_{1j}) \sum_{i=1}^{n} \log \phi(y_{ij}|\alpha_{2i}, \varsigma_{2i}^2) \Big\}.
\end{aligned}$$

At a given iterate of the parameters, $\boldsymbol{\theta}^{(n)}$, the E-step involves taking expectation with respect to the density $f(\mathbf{w}, \boldsymbol{\gamma}|\mathbf{y}, \boldsymbol{\theta}^{(n)}) = f(\mathbf{w}|\boldsymbol{\gamma}, \mathbf{y}, \boldsymbol{\theta}^{(n)}) f(\boldsymbol{\gamma}|\mathbf{y}, \boldsymbol{\theta}^{(n)})$. Given the importance indicators, $\boldsymbol{\gamma}$, the posterior probability that a gene methylates in cluster c is

$$\begin{aligned}
E[w_{cj}|\mathbf{y}, \boldsymbol{\gamma}] = & \gamma_j \frac{\pi_{1c}^{(n)} \prod_{i\in c} \phi(y_{ij}|\mu_{1i}^{(n)}, \sigma_{1i}^{2(n)})}{\pi_{1c}^{(n)} \prod\limits_{i\in c} \phi(y_{ij}|\mu_{1i}^{(n)}, \sigma_{1i}^{2(n)}) + \pi_{2c}^{(n)} \prod\limits_{i\in c} \phi(y_{ij}|\mu_{2i}^{(n)}, \sigma_{2i}^{2(n)})} \\
& + (1-\gamma_j) \frac{\pi_1^{(n)} \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{1i}^{(n)}, \varsigma_{1i}^{2(n)})}{\pi_1^{(n)} \prod\limits_{i=1}^{n} \phi(y_{ij}|\alpha_{1i}^{(n)}, \varsigma_{1i}^{2(n)}) + \pi_2^{(n)} \prod\limits_{i=1}^{n} \phi(y_{ij}|\alpha_{2i}^{(n)}, \varsigma_{2i}^{2(n)})}.
\end{aligned} \tag{5}$$

Note that the E-step involves replacing w_{cj} with (5) in the complete data loglikelihood above, which makes it linear in γ_j, and then taking expectation with respect to the posterior distribuiton of $\boldsymbol{\gamma}$. But the posterior probability that gene j discriminates well is simply given by

$$E[\gamma_j|\mathbf{y}] = \frac{p^{(n)} f_1^{(n)}(\mathbf{y}_j)}{p^{(n)} f_1^{(n)}(\mathbf{y}_j) + (1-p^{(n)}) f_2^{(n)}(\mathbf{y}_j)}, \tag{6}$$

where $f_1^{(n)}(\mathbf{y}_j) = \prod_c \left(\pi_{1c}^{(n)} \prod_{i\in c} \phi(y_{ij}|\mu_{1i}^{(n)}, \sigma_{1i}^{2(n)}) + \pi_{2c}^{(n)} \prod_{i\in c} \phi(y_{ij}|\mu_{2i}^{(n)}, \sigma_{2i}^{2(n)}) \right)$ and $f_2^{(n)}(\mathbf{y}_j) = \pi_1^{(n)} \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{1i}^{(n)}, \varsigma_{1i}^{2(n)}) + \pi_2^{(n)} \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{2i}^{(n)}, \varsigma_{2i}^{2(n)})$. The fact that the Q-function has a closed form, through the use of the tower property of conditional expectation, leads to closed form solutions to the M-step.

3 Application to the Erasmus Data

We compare our clustering results to that of Figueroa et al. (2010) and also summarize our classification results on a subset of patients that have

known subtypes of AML.

3.1 Clustering Results

In Figueroa et al. (2010) a cut-off of $K = 16$ clusters was selected subjectively, although values ranging from $K = 13, \ldots, 21$ also made biological sense. We focused on the same set of genes (those with standard deviations > 1) to get a direct comparison. Our hierarchical likelihood based algorithm resulted in a partition with maximum likelihood at 16 clusters as well. A graphical comparison of the two methods is displayed in Figure 1. The two methods both did very well in identifying known and well characterized subtypes of AML, but there were some discrepancies between patients with novel and not well characterized subtypes.

3.2 Classification Results

In the Erasmus data set we looked at 62 patients with 3 well known subtypes of AML. We randomly split the patients into a training set and a test set making sure that the same percentage of patients from each group was present in the training set. We then fit the model on the training set and classified each of the patients in the test set according to the classification rule given in subsection 1.2. This was done for percentage values (of patients in the training set) ranging from $70 - 90\%$ and repeated $1,000$ times. Summary of the classification success rates is given in Table 1.

3.3 Discussion

Further research is needed for developing the clustering algorithm. For example, it is not clear what is the best method for selecting the subset of discriminating genes, J_d, before running the algorithm. It also seems reasonable to apply some kind of stochastic search on the partition space, as in Booth et al. (2008) But overall, the likelihood based model gives some very promising results in clustering and classifying patients with AML. Our method, through the posterior expectations of our EM algorithm, automatically provides us with information about which genes methylate differently across clusters. Thus there is no need to perform ANOVA on each gene after coming up with a candidate partition. Moreover we learn exactly which genes methylate and which don't in all clusters. This could become valuable information for understanding the biological bases of different subtypes of AML.

References

Booth, J.G., Casella, G., and Hobert, J.P. (2008). *Clustering using objective functions and stochastic search.* Journal of the Royal Statistical Society, Series B, **70**.

TABLE 1. Classification success rates (1,000 reps each).

Percentage in training set	Success rate
70%	93.9%
80%	94.8%
90%	95.1%

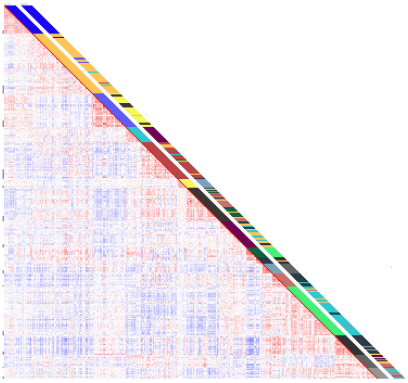


FIGURE 1. A correlation heat map for the 344 patients. The first diagonal strip represents the clustering results from Figueroa et al. (2010) and the second diagonal strip our results.

Figueroa, M.E., Lugthart, S., et al. (2010). *Epigenetic Signatures Identify Biologically Distinct Subtypes in Acute Myeloid Leukemia.* Cancer Cell.

Tadesse, M.G., Sha, N., and Vannucci, M. (2005). *Bayesian Variable Selection in Clustering High-Dimensional Data.* Journal of the American Statistical Association.

Generalization of the Order-Restricted Information Criterion: Illustrated

Rebecca M. Kuiper[1], Herbert Hoijtink[1]

[1] Department of Methodology and Statistics, Utrecht University. *Address for correspondence*: PO Box 80140, 3508 TC Utrecht, the Netherlands. E-mail: R.M.Kuiper@uu.nl.

Abstract: Often researchers have a hypothesis about the parameters of interest - say μ_1, μ_2, and μ_3 - with a certain direction (e.g., $\mu_1 > \mu_2 > \mu_3$). However, they test this hypothesis by testing $H_0 : \mu_1 = \mu_2 = \mu_3$ with an F test. The Order-Restricted Information Criterion (ORIC) of Anraku (1999) can be used to evaluate this type of hypotheses directly. However, this ORIC can only evaluate simple order restrictions. Therefore, we will propose and illustrate a generalization of the ORIC, which can handle a more general form of order restrictions.

Keywords: Model selection; Analysis of variance; Akaike information criterion; order-restricted information criterion; Interaction effects.

1 Introduction

Most researchers are able to specify reasonable "order-restricted" hypotheses - say $H_1 : \mu_1 > \mu_2 > \mu_3$ -, since they are expert in their research field. In that case, the researcher should evaluate this order-restricted hypothesis H_1 by using a confirmatory method. For more details and comparisons of confirmatory techniques see Kuiper and Hoijtink (2010). One confirmatory model selection technique is the Order-Restricted Information Criterion (ORIC) of Anraku (1999). The ORIC is a generalization of the Akaike information criterion (AIC) (Akaike, 1974), which can handle simple order restrictions, which are of the form $\mu_1 \geq \ldots \geq \mu_k$. However, this ORIC can only evaluate simple order restrictions. The generalization of the ORIC (genORIC) can be used for all order restrictions of the type $R\mu \leq r$, with μ a vector of length k, r a vector of length c_m, and R a $c_m \times k$ matrix. We will not show the derivation of genORIC in this paper, but we will describe the genORIC in the context of the ANOVA model in the next section. Subsequently, we illustrate the genORIC with use of the research of Berzonky, Kleven, and Leach (2003).

2 The genORIC

We will look (like Anraku (1999)) at the case where

$$y_{ij} \sim \mathcal{N}(\mu_i, \tau_i \sigma^2), \text{ for } i = 1, \ldots, k \text{ and } j = 1, \ldots, n_i$$

where τ_i is known and σ^2 unknown.
It can be shown that the genORIC, like the ORIC, is calculated by

$$genORIC_m \quad = \quad -2 \ \log L_m(\tilde{\mu}_m, \tilde{\sigma}^2_m | y) + 2 \ PT_m,$$

where $\log L_m$ is the log likelihood for hypothesis H_m, $\tilde{\mu}_m = (\tilde{\mu}_{m1}, \ldots, \tilde{\mu}_{mk})$ are the restricted means, that is, the values for which the likelihood is maximized subject to the restrictions in hypothesis H_m, and PT_m is the penalty term for hypothesis H_m, which can be seen as the expected number of distinct mean values plus 1 (because of the unknown variance term):

$$PT_m = 1 + \sum_{l=1}^{k} w_{ml}(k, V^*, \mathcal{C}_m) \cdot l,$$

with $w_{ml}(k, V^*, \mathcal{C}_m)$ the level probability for hypothesis H_m, where $V^* = \sigma^2 \ \text{diag} \left\{ \frac{\tau_1}{n_1}, \ldots, \frac{\tau_k}{n_k} \right\}$. A level probability is the probability that there are l distinct mean values (i.e., levels) among the k restricted means in Hypothesis m. More details can be found in Silvapulle and Sen (2005) and Robertson, Wright, and Dykstra (1988).

3 Illustration of genORIC

In short, Berzonky et al. (2003) looked at the effects of parthenogenesis on wheat embryo formation with and without maize pollination. The dependent variable is embryo formation frequency (EFF). In their research, they performed a two-way ANOVA. The two factors are genotype and treatment. The genotype factor consists of 4 groups: Salmon(K), Salmon, CS(K), and CS. The factor treatment distinguishes 2 groups, namely the group with (p) and the group without (np) maize pollination. In total there are 8 groups. Each group consists of 15 plants, that is, $n_i = 15$ for $i = 1, \ldots, 8$. More details can be found in Berzonky et al. (2003).
Let $\mu_{G,T}$ be the mean of EFF for genotype G and treatment T. Berzonky et al. (2003) expect that

$$\mu_{Salmon(K),T} \quad > \quad \mu_{G',T}, \tag{1}$$

$$\mu_{G,p} \quad > \quad \mu_{G,np}, \tag{2}$$

$$\mu_{Salmon(K),p} - \mu_{Salmon(K),np} \quad > \quad \mu_{G',p} - \mu_{G',np}. \tag{3}$$

for G' = Salmon, CS(K), and CS and $T = p, np$. Since the expected interactions in (3) cannot be written in terms of simple order restrictions, we need to use of genORIC.
It depends on the researchers' or other researchers' expectations which hypotheses are also evaluated with the genORIC. For example, there could be other authors who expect (instead of (3)) that

$$\mu_{Salmon(K),p} - \mu_{Salmon(K),np} \quad > \quad 2\ (\mu_{G',p} - \mu_{G',np}). \tag{4}$$

In that case, the researcher can define a competing hypothesis

$$H_2: \qquad (1), (2), \text{ and } (4).$$

Let μ_1, μ_2, μ_3, μ_4, μ_5, μ_6, μ_7, and μ_8 represent $\mu_{Salmon(K),p}$, $\mu_{Salmon,p}$, $\mu_{CS(K),p}$, $\mu_{CS,p}$, $\mu_{Salmon(K),np}$, $\mu_{Salmon,np}$, $\mu_{CS(K),np}$, and $\mu_{CS(K),np}$, respectively.
To ensure that not a weak hypothesis is chosen, the unconstrained hypothesis (H_u) should be included (Kuiper and Hoijtink, 2010) and although the null hypothesis (H_0) should only be included when there is real interest in H_0 (Kuiper and Hoijtink, 2010), we include it for illustration purposes, where

$$\begin{aligned} H_u: &\qquad \mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6, \mu_7, \mu_8, \\ H_0: &\qquad \mu_1 = \mu_2 = \mu_3 = \mu_4, \mu_5 = \mu_6 = \mu_7 = \mu_8. \end{aligned}$$

We will evaluate these four hypotheses with the genORIC. Since the original data of Berzonky et al. (2003) was not available, we generated data from a normal distribution with means and standard deviations (Table 1) of the non-transformed data reported in Berzonky et al. (2003).
As for other information criteria, the hypothesis with the smallest value for the genORIC is the preferred hypothesis. The resulting values for the likelihood, penalty, and genORIC for these hypotheses are given in Table 2. It can be concluded that hypothesis H_1, that is, the hypothesis based on Berzonky et al. (2003), is the preferred hypothesis.

4 Discussion

With the ORIC derived by Anraku (1999) we cannot evaluate hypotheses with interaction effects. We can only evaluate simple order restrictions: $\mu_1 \geq \ldots \geq \mu_k$. The genORIC can be used for all order restrictions of the type $\mu \in \mathcal{C}_m$. A special case of $\mu \in \mathcal{C}_m$ is $R\mu \leq r$, with μ a vector of length k, r a vector of length c_m, and R a $c_m \times k$ matrix. Note that the simple order restrictions are a special case of $\mathcal{C}_m$. Thus, the ORIC is a special case of the genORIC. Software regarding the ORIC (called "Confirmatory ANOVA" and "Comparison Of Means") and the genORIC can be found at `www.fss.uu.nl/ms/Kuiper`.

TABLE 1. *Number of Observations (n_i), Sample Means ($\overline{y}_i$), and Sample Standard Deviations (sd_i) for the Four Genotypes and the Two Treatments*

maize pollination					no maize pollination				
i	Genotype	n_i	$\overline{y}_i$	sd_i	i	Genotype	n_i	$\overline{y}_i$	sd_i
1	S(K)	15	32.00	4.31	5	S(K)	15	14.00	4.22
2	S	15	21.00	4.47	6	S	15	0.01	0.01
3	CS(K)	15	7.00	4.05	7	CS(K)	15	0.10	0.09
4	CS	15	14.00	6.03	8	CS	15	0.22	0.21

TABLE 2. *The genORIC of the Specified Hypotheses (H_m)*

m	H_m	PT_m	$\log L_m$	$genORIC_m$
0	H_0	2.00	-462.32	928.65
1	H_1	5.47	-328.30	**667.54**
2	H_2	5.30	-353.32	717.23
u	H_u	9.00	-327.09	672.17

Acknowledgments: This study has been funded by the Netherlands Organization for Scientific Research NWO-VICI-453-05-002.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control*, **19**, 716-723.

Anraku, K. (1999). An Information Criterion for Parameters under a Simple Order Rrestriction. *Biometrika*, **86**, 141-152.

Berzonsky, W.A., Kleven, S.L. and Leach. G.D. (2003). The effects of parthenogenesis on wheat embryo formation and haploid production with and without maize pollination. *Euphytica*, **133**, 285-290.

Kuiper, R. M. and Hoijtink, H. (2010). Comparisons of Means Using Exploratory and Confirmatory Approaches. *Psychological Methods*, **15**, 69-86.

Robertson, T., Wright, F.T. and Dykstra, R.L. (1988). *Order restricted statistical inference*. Chichester: Wiley.

Silvapulle, M.J. and Sen, P.K. (2005). *Constrained Statistical Inference*. New Jersey: Wiley.

Additive model for the conditional location and dispersion of a smooth distribution when the observed data are interval censored

Philippe Lambert[1,2]

[1] Institut des sciences humaines, Univ. de Liège, Belgium (p.lambert@ulg.ac.be)
[2] Institut de statistique, biostatistique et sciences actuarielles (IBSA), Université catholique de Louvain, Belgium.

Abstract: An additive model for the location and dispersion of a continuous response with an arbitrary smooth conditional distribution is proposed. B-splines are used to specify the three components of the model. It can be extended to deal with interval censored data and multiple covariates. As an illustration, the relation between age, the number of years of full-time education and total net income (provided as intervals) of isolated persons is studied from household survey data.

Keywords: Interval censored data ; additive model ; Smooth distribution.

1 Smooth specification of the conditional distribution

For simplicity, assume that one wants to specify how the conditional distribution of a single continuous random variable Y changes with a single continuous covariate X. Denote by $f_{Y|X}(y|x)$ the corresponding conditional density.
Further assume that conditionally on $X = x$, the distribution of $Z = (Y-\mu_x)/\sigma_x$ does not depend on x for suitably specified conditional location and dispersion parameters, μ_x and σ_x. One can find $a > 0$ such that the support of Z is most likely included in $(-a, a)$. For example, if μ_x and σ_x^2 denote the first two conditional moments, then, by Chebyshev's theorem, one knows that $\Pr(Z \leq a) \geq 1 - a^{-2}$. In most practical situations, that probability is far above that lower bound. Therefore, taking $a = 6$ (say) usually provides satisfactory results.
Consider now a cubic B-splines basis $\{b_k(\cdot) : k = 1, \ldots, K\}$ associated to 20 (say) equidistant knots on $(-a, a)$. Let $\{\mathcal{J}_j : j = 1, \ldots, J\}$ be a partition of $(-a, a)$ into a large number (100, say) of consecutive bins of equal width Δ with midpoints $u_{j=1}^J$. Then, the probability to observe Z in $\mathcal{J}_j$ is specified as

$$\int_{\mathcal{J}_j} f_Z(z)dz = \pi_j = \frac{\exp(\sum_k b_k(u_j)\phi_k^*)}{\sum_{l=1}^L \exp\left(\sum_k b_k(u_l)\phi_k^*\right)} \approx f_Z(u_j)\Delta \qquad (1)$$

where $\phi^* = (\phi_1^*, \ldots, \phi_K^*)'$ is a vector of spline parameters. As $\pi_j(\phi^* + c) = \pi_j(\phi^*)$ for any scalar c, one should constrain ϕ^* for identifiability. We suggest to work with ϕ,

$$\phi_k = \phi_k^* - \log\left(\sum_{l=1}^{L} \exp\left(\sum_k b_k(u_l)\phi_k^*\right)\right),$$

such that

$$\pi_j = \exp\left(\sum_k b_k(u_j)\phi_k\right).$$

2 Smooth model for location and dispersion

Assume for simplicity that a single continuous covariate X takes values in $(0, 1)$ after relocation and rescaling. Provided that these are smooth, the functional forms of

$$\mu_x = \mathrm{E}(Y|X = x) \quad \text{and} \quad \log\sigma_x = .5\log\mathrm{V}(Y|X = x),$$

can be approximated using a linear combination of the elements of a (large) B-splines basis $\{b_k^X(\cdot) : k = 1, \ldots, K_X\}$ on $(0, 1)$:

$$\mu_x = \mu(x|\lambda) = \sum_{k=1}^{K_X} b_k^X(x)\lambda_k \quad ; \quad \log\sigma_x = \log\sigma(x|\delta) = \sum_{k=1}^{K_X} b_k^X(x)\delta_k.$$

3 Penalties and Bayesian formulation

3.1 Roughness penalty

The flexibility provided by the large numbers of B-splines, K and K^X, can be counterbalanced by a roughness penalty in a frequentist setting (Eilers and Marx, 1996) or a suitable prior in a Bayesian framework (Lang and Brezger, 2004). The chosen order for the penalty will depend on the desired limiting behavior for the functionals for large values of the penalty. We suggest to work with

- a 3rd order penalty for the conditional distribution, yielding a normal distribution at the limit:

$$\mathrm{pen}_\phi = -0.5\tau_\phi \sum_k (\phi_k - 3\phi_{k-1} + 3\phi_{k-2} - \phi_{k-3})^2 = -0.5\tau_\phi \phi' P_\phi \phi \quad ;$$

- a 2nd order penalty for the conditional location, yielding a linear model at the limit:

$$\mathrm{pen}_\lambda = -0.5\tau_\lambda \sum_k (\lambda_k - 2\lambda_{k-1} + \lambda_{k-2})^2 = -0.5\tau_\lambda \lambda' P_\lambda \lambda \quad ;$$

- a 1st order penalty for the conditional dispersion, yielding an homoskedastic model at the limit:

$$\text{pen}_\delta = -0.5\tau_\delta \sum_k (\delta_k - \delta_{k-1})^2 = -0.5\tau_\delta \delta' P_\delta \delta.$$

In Bayesian terms, it translates into prior distributions on the spline coefficients:

$$p(\phi|\tau_\phi) \propto \tau_\phi^{\rho(P_\phi)/2} \exp\left(-0.5\tau_\phi\ \phi' P_\phi \phi\right), \tag{2}$$

$$p(\lambda|\tau_\lambda) \propto \tau_\lambda^{\rho(P_\lambda)/2} \exp\left(-0.5\tau_\lambda\ \lambda' P_\lambda \lambda\right), \tag{3}$$

$$p(\delta|\tau_\delta) \propto \tau_\delta^{\rho(P_\delta)/2} \exp\left(-0.5\tau_\delta\ \delta' P_\delta \delta\right). \tag{4}$$

3.2 Likelihood and identification penalty

Given $\theta = (\phi', \lambda', \delta')'$, one can associate a quantity z_i to each observation (x_i, y_i) such that $z_i = (y_i - \mu_{x_i})/\sigma_{x_i}$. If one denotes by $m_j = m_j(\{x_i, y_i\}, \lambda', \delta')$ $(j = 1, \ldots, J)$, the number of z_i's belonging to bin $\mathcal{J}_j$, then the conditional joint distribution of $(M_1, \ldots, M_J)$ is multinomial,

$$(M_1, \ldots, M_J|\theta) \sim \text{Mult}(n; \pi_1, \ldots, \pi_J),$$

for values of π_j given by Eq. (1). Therefore, the log-likelihood will be

$$\log L(\theta|\text{data}) = \sum_{j=1}^{J} m_j \log \pi_j.$$

An extra penalty should be added to force the desired interpretation for the location and dispersion parameters. For example, to interpret μ_x and σ_x as the conditional mean and standard deviation, one should make sure that the mean and variance of Z are 0 and 1, respectively. Using Eq. (1), one has

$$\text{E}(Z) \approx \bar{z} = \frac{1}{J}\sum_{j=1}^{J} u_j \pi_j \quad ; \quad \text{V}(Z) \approx s_z^2 = \frac{1}{J}\sum_{j=1}^{J} u_j^2 \pi_j - \bar{z}^2,$$

where u_j is the midpoint of bin $\mathcal{J}_j$. Therefore, one could add a large identifiability penalty

$$\text{pen}_{id} = -\kappa \left\{\bar{z}^2 + (s_z^2 - 1)^2)\right\},$$

to force mean 0 and variance 1 for the distribution of the standardized responses. This suggests working with the penalized log-likelihood

$$\log L_{\text{pen}}(\theta|\text{data}) = \sum_{j=1}^{J} m_j \log \pi_j + \text{pen}_{id} \tag{5}$$

In a Bayesian framework, the log of the joint posterior is simply the sum of (5) and of the logarithms of (2),(3) and (4). That expression can be used with a Metropolis-within-Gibbs algorithm to sample the joint posterior. From the generated chain, one can build point estimates and credible regions for the spline parameters and any derived quantity.
Starting values for the chain can be obtained by taking values of ϕ corresponding to a standard normal distribution and values of λ, δ corresponding to the fit of an additive model for the mean with τ_λ selected to minimize BIC in an homoskedastic setting.

4 Extension to interval censored data

Assume now that the response data take the form of intervals $\{(x_i, (y_i^L, y_i^U)\}$. Then, the standardized intervals are (z_i^L, z_i^U) where

$$z_i^L = \frac{y_i^L - \mu_{x_i}}{\sigma_{x_i}} \quad ; \quad , z_i^U = \frac{y_i^U - \mu_{x_i}}{\sigma_{x_i}}.$$

If c_{ij} is the proportion of bin $\mathcal{J}_j$ contained in (z_i^L, z_i^U), then the penalized log-likelihood that one had in Eq. (5) for precisely observed responses becomes

$$\log L_{\text{pen}}(\theta|\text{data}) = \sum_{i=1}^{I} \log \left(\sum_{j=1}^{J} c_{ij} \pi_j \right) + \text{pen}_{id}.$$

The derivation of the joint posterior and the strategy for exploring it are unchanged.

5 Illustration on simulated data

One thousand data $\{y_i : i = 1, 1000\}$ were simulated from a distribution with conditional mean $\mu_x = \sin(2\pi x)$, conditional standard deviation $\sigma_x = \exp\{6\cos(4\pi x)\}$ and a skewed density; see the solid lines in Fig. 1. The points on the left panel of Fig. 1 are the midpoints of the observed interval data $(y_i - u_i^L \sigma_{x_i}, y_i + u_i^U \sigma_{x_i})$ where u_i^L, u_i^U are i.i.d. uniform on (0,1). One can see from the fitted functionals (dashed lines) and from the 95% credible regions that the 3 components of the model are pretty well estimated. Of course, in practice, if the available information is too sparse to provide fine estimates of some of the components, the penalties will guide us, at the limit, towards a classical linear regression setting with a normal pivotal distribution.

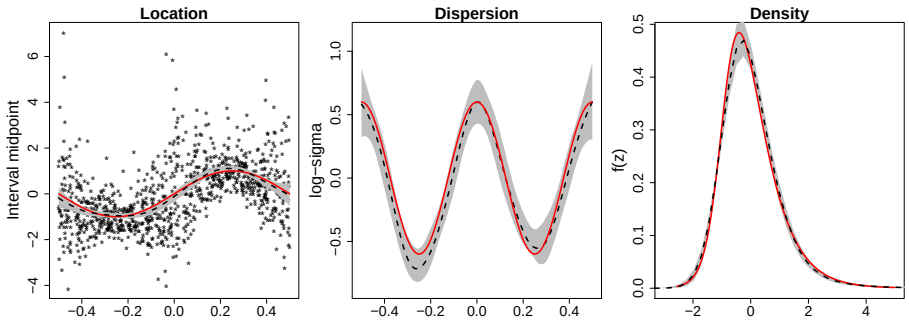


FIGURE 1. Simulated data: theoretical (solid) and fitted (dashed) conditional mean, log standard deviation and density ; the grey areas correspond to pointwise 95% credible intervals.

6 Application

The data of interest were collected in 2002 through the European Social Survey (2002). Here we focus on the net monthly income of persons living alone in Belgium ($n = 213$) and its relation with age (57.8 ± 19.2) and the number of years of full-time education completed (11.4 ± 3.61). Income (in euros) were reported in one of the following 12 intervals: 1: < 150 2: $[150, 300[$, 3: $[300, 500[$, 4: $[500, 1.000[$, 5: $[1.000, 1.500[$, 6: $[1.500, 2.000[$, 7: $[2.000, 2.500[$, 8: $[2.500, 3.000[$, 9: $[3.000, 5.000[$, 10: $[5.000, 7.500[$, 11: $[7.500, 10.000[$, 12: ≥ 10.000.
The `age` and `educ` components in the additive models for the conditional mean and (log-) standard deviation of the net monthly income are reported in Fig. 2. It suggests a significant increase (at a given age) of the mean and standard deviation of the net income with the number of years of education. An effect of age on the mean income is also visible and most likely associated to people losing their job in their late fifties or retiring at 65 (or a bit earlier).

Acknowledgments: Financial support from the FSR research grant nr FSRC-08/42 from the University of Liège is gratefully acknowledged.

References

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

European Social Survey Round 1 Data (2002). Data file edition 6.1. Nor-

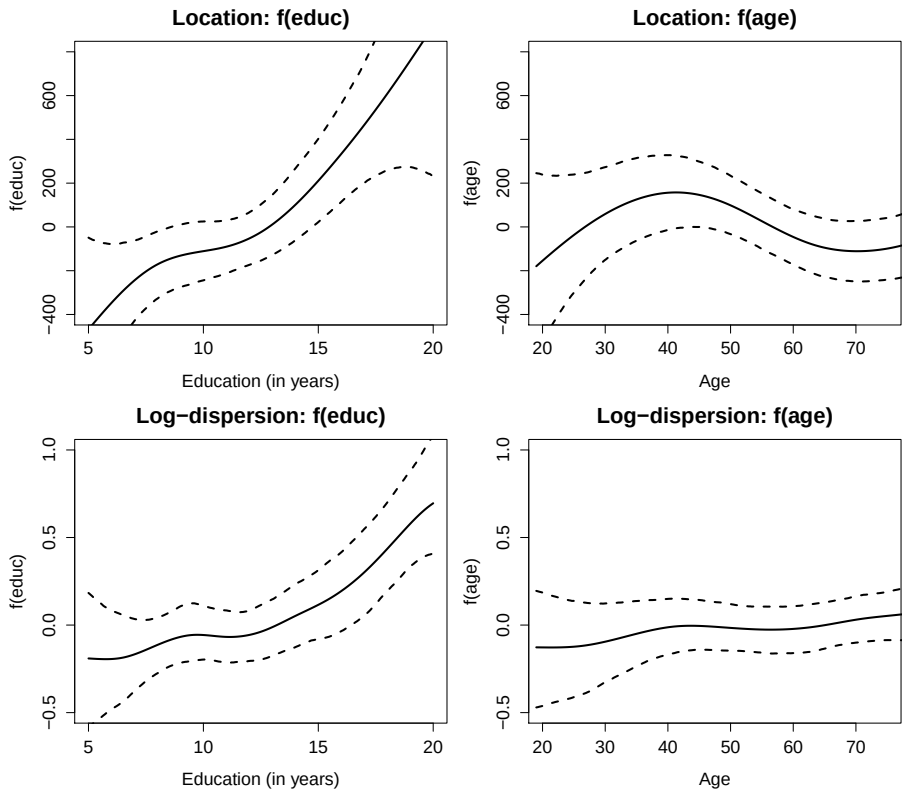


FIGURE 2. Income data: `age` and `educ` components (solid line) in the additive models for the mean and for the log-standard deviation ; dashed lines delimit pointwise 95% credible intervals.

wegian Social Science Data Services, Norway, Data Archive and distributor of ESS data.

Lambert, P. and Eilers, P. H. C. (2009) Bayesian density estimation from grouped continuous data. *Computational Statistics and Data Analysis*, **53**: 1388-1399.

Lambert, P. (2009) Smooth semi- and nonparametric Bayesian estimation of bivariate densities from bivariate histogram data. *Technical Report* TR09038, Interuniversity Attraction Pole, UCL, Belgium.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, **13**, 183–212.

Hierarchical Structured Additive Regression

Stefan Lang[1], Nikolaus Umlauf[1], Wolfgang Brunauer[2]

[1] Department of Statistics , Universität Innsbruck, Universitätsstraße 15, A-6020 Innsbruck, Tel: +43 512 507 - 7101, email: *stefan.lang@uibk.ac.at*
[2] Immobilien Rating GmbH, Taborstr. 1–3, A-1020 Wien

Abstract: Models with structured additive predictor (STAR) provide a very broad and rich framework for complex regression modeling. They can deal simultaneously with nonlinear covariate effects and time trends, unit- or cluster specific heterogeneity, spatial heterogeneity and complex interactions between covariates of different type. We present a hierarchical version of STAR models, i.e. the regression coefficients of a particular nonlinear term may obey another regression model with structured additive predictor. The proposed model may be regarded as a an extended version of a multilevel model with nonlinear covariate terms in every level of the hierarchy. The hierarchical structure of the model is also utilized for efficient Markov chain Monte Carlo (MCMC) inference in a fully Bayesian approach.

Keywords: MCMC; multilevel models; P-splines; spatial smoothing

1 Structured Additive Regression Models

Suppose we are given n observations $(y_i, \mathbf{z}_i)$ of a continuous response variable y and a vector of covariates $\mathbf{z} = (z_1, \ldots, z_q)$. STAR models (Fahrmeir et al. 2004, Belitz and Lang 2008) are given by

$$y_i = \eta_i + \varepsilon_i = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \varepsilon_i, \tag{1}$$

with $\varepsilon_i \sim N(0, \sigma^2)$ i.i.d. and possibly nonlinear functions f_j. The functions f_j are rather general and may comprise for instance one or two dimensional P-splines (Eilers and Marx 1996) for modeling the effect of continuous covariates, Markov-random fields, Gaussian random fields (kriging) and Gaussian random effects for modeling spatial- or cluster effects, etc.
The nonlinear functions in (1) are modeled by a basis function approach, i.e. a particular function f of z may be decomposed in the form

$$f(z) = \sum_{k=1}^{K} \beta_k B_k(z), \tag{2}$$

where the B_k's are known basis or indicator functions that are scaled with the regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$. Defining the $n \times K$ design

matrix $\mathbf{Z}$, which contains in each column k the evaluations of basis function B_k, i.e. $\mathbf{Z}[i,k] = B_k(z_i)$, the vector $\mathbf{f} = (f(z_1), \ldots, f(z_n))'$ can be written in matrix notation as $\mathbf{Z}\boldsymbol{\beta}$. Accordingly, the additive predictor in (1) can be written as

$$\boldsymbol{\eta} = \mathbf{Z}_1\boldsymbol{\beta}_1 + \ldots + \mathbf{Z}_q\boldsymbol{\beta}_q \tag{3}$$

In this paper we propose a hierarchical or multilevel version of STAR models. That is the regression coefficients $\boldsymbol{\beta}_j$ of a term f_j may themselves obey a regression model with structured additive predictor, i.e.

$$\boldsymbol{\beta}_j = \boldsymbol{\eta}_j + \boldsymbol{\varepsilon}_j = \mathbf{Z}_{j1}\boldsymbol{\beta}_{j1} + \ldots + \mathbf{Z}_{jq_j}\boldsymbol{\beta}_{jq_j} + \boldsymbol{\varepsilon}_j, \tag{4}$$

where the terms $\mathbf{Z}_{j1}\boldsymbol{\beta}_{j1}, \ldots, \mathbf{Z}_{jq_j}\boldsymbol{\beta}_{jq_j}$ correspond to additional nonlinear functions $f_{j1}, \ldots, f_{jq_j}$ and $\boldsymbol{\varepsilon}_j \sim N(\mathbf{0}, \tau_j^2\mathbf{I})$ is a vector of i.i.d. Gaussian errors. A third or even higher levels in the hierarchy are possible by assuming that the second level regression parameters $\boldsymbol{\beta}_{jl}$, $l = 1, \ldots, q_j$, obey again a STAR model. In that sense, the model is composed of a hierarchy of complex structured additive regression models.

2 Structure of the priors

We distinguish two types of priors: "direct" or "basic" priors for the regression coefficients $\boldsymbol{\beta}_j$ (or $\boldsymbol{\beta}_{jl}$ in a second level equation) and compound priors (4).

2.1 General form of basic priors

The general form of the basic prior is given by a (possibly improper) multivariate Gaussian density

$$p(\boldsymbol{\beta}|\tau^2) \propto \left(\frac{1}{\tau^2}\right)^{rk(\mathbf{K})/2} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\beta}'\mathbf{K}\boldsymbol{\beta}\right) \cdot I(\mathbf{A}\boldsymbol{\beta} = \mathbf{0}), \tag{5}$$

where $I(\cdot)$ is the indicator function. The key components of the prior are the penalty matrix $\mathbf{K}$, the variance parameter τ^2 and the constraint $\mathbf{A}\boldsymbol{\beta} = \mathbf{0}$. The structure of the penalty or prior precision matrix $\mathbf{K}$ depends on the covariate type and on our prior assumptions about smoothness of f. The amount of smoothness is governed by the variance parameter τ^2. A conjugate inverse Gamma prior is employed for τ^2 (as well as for the overall variance parameter σ^2), i.e. $\tau^2 \sim IG(a, b)$ with small values such as $a = b = 0.001$ for the hyperparameters a and b resulting in an uninformative prior on the log scale. The term $I(\mathbf{A}\boldsymbol{\beta} = \mathbf{0})$ imposes required identifiability constraints on the parameter vector. Specific examples for modeling nonlinear terms are one or two dimensional P-splines for nonlinear effects of continuous covariates, or Gaussian Markov random fields and Gaussian fields (kriging) for modeling spatial heterogeneity, see Lang et al. (2010) for details.

2.2 Compound priors

In the vast majority of cases a compound prior is used if a covariate $z_j \in \{1, \ldots, K\}$ is a unit- or cluster index and z_{ij} indicates the cluster observation i pertains to. Then the design matrix $\mathbf{Z}_j$ is a $n \times K$ incidence matrix with $\mathbf{Z}_j[i,k] = 1$ if the i-th observation belongs to cluster k and zero else. The $K \times 1$ parameter vector $\boldsymbol{\beta}_j$ is the vector of regression parameters, i.e. the k-th element in $\boldsymbol{\beta}_j$ corresponds to the regression coefficient of the k-th cluster. Using the compound prior (4) we obtain an additive decomposition of the cluster specific effect. The covariates z_{jl}, $l = 1, \ldots, q_j$, in (4) are cluster specific covariates with possible nonlinear cluster effect. By allowing a full STAR predictor (as in the level-1 equation) a rather complex decomposition of the cluster effect $\boldsymbol{\beta}_j$ including interactions is possible. A special case arises if cluster specific covariates are not available. Then the prior for $\boldsymbol{\beta}_j$ collapses to $\boldsymbol{\beta}_j = \boldsymbol{\varepsilon}_j \sim N(\tau_j^2 \mathbf{I})$ and we obtain a simple i.i.d. Gaussian cluster specific random effect with variance parameter τ_j^2.
Another special situation arises if the data are grouped according to some discrete geographical grid and the cluster index z_{ij} denotes the geographical region observation i pertains to. For instance, in our applications on hedonic house price modeling in section 4 for every observation the municipality, district and county the house is located is given. Then the compound prior (4) models a complex spatial heterogeneity effect with possibly nonlinear effects of region specific covariates z_{jl}.
In a number of applications geographical information and spatial covariates are given at different resolutions. For instance, in our case study on hedonic house prices, the municipalities (level-2) are nested within districts (level-3), which are nested in counties (level-4). This allows to model a spatial effect over three levels of the form

$$\begin{aligned}
\boldsymbol{\beta}_j &= \mathbf{Z}_{j1}\boldsymbol{\beta}_{j1} + \mathbf{Z}_{j2}\boldsymbol{\beta}_{j2} + \ldots + \boldsymbol{\varepsilon}_j, \\
\boldsymbol{\beta}_{j1} &= \mathbf{Z}_{j11}\boldsymbol{\beta}_{j11} + \mathbf{Z}_{j12}\boldsymbol{\beta}_{j12} + \ldots + \boldsymbol{\varepsilon}_{j1}. \\
\boldsymbol{\beta}_{j11} &= \mathbf{Z}_{j111}\boldsymbol{\beta}_{j111} + \mathbf{Z}_{j112}\boldsymbol{\beta}_{j112} + \ldots + \boldsymbol{\varepsilon}_{j11}.
\end{aligned}$$

Here, the first covariate z_{j1} in the municipality specific effect is another cluster indicator that indicates the district in which the municipalities are nested. Hence $\mathbf{Z}_{j1}$ is another incidence matrix and $\boldsymbol{\beta}_{j1}$ is the vector of district specific effects modeled through the level-3 equation. In the level-3 equation covariate z_{j11} is again a cluster indicator for the county. The county specific effect is given in the level-4 equation.
Other possibilities for compound priors can be found in Lang et al. (2010).

3 Sketch on MCMC Inference

For the sake of simplicity we restrict the presentation to a two level hierarchical model with one level-2 equation for the regression coefficients of the

first term $\mathbf{Z}_1\boldsymbol{\beta}_1$. That is, the level-1 equation is $\mathbf{y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon}$ with predictor (1) and errors $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. The level-2 equation is of the form (4) with $j = 1$.

The full conditionals for the regression coefficients $\boldsymbol{\beta}_1$ with the compound prior (4) and the coefficients $\boldsymbol{\beta}_j$, $j = 2, \ldots, q$, $\boldsymbol{\beta}_{1l}$, $l = 1, \ldots, q_1$ with the basic prior (5) are all multivariate Gaussian. The respective posterior precision $\boldsymbol{\Sigma}^{-1}$ and mean $\boldsymbol{\mu}$ is given by

$$
\begin{aligned}
\boldsymbol{\Sigma}^{-1} &= \tfrac{1}{\sigma^2}\left(\mathbf{Z}_1'\mathbf{Z}_1 + \tfrac{\sigma^2}{\tau_1^2}\mathbf{I}\right), & \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} &= \tfrac{1}{\sigma^2}\mathbf{Z}_1'\mathbf{r} + \tfrac{1}{\tau_1^2}\boldsymbol{\eta}_1, & &(\boldsymbol{\beta}_1),\\
\boldsymbol{\Sigma}^{-1} &= \tfrac{1}{\sigma^2}\left(\mathbf{Z}_j'\mathbf{Z}_j + \tfrac{\sigma^2}{\tau_j^2}\mathbf{K}_j\right), & \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} &= \tfrac{1}{\sigma^2}\mathbf{Z}_j'\,\mathbf{r}, & &(\boldsymbol{\beta}_j), \qquad (6)\\
\boldsymbol{\Sigma}^{-1} &= \tfrac{1}{\tau_1^2}\left(\mathbf{Z}_{1l}'\mathbf{Z}_{1l} + \tfrac{\tau_1^2}{\tau_{1l}^2}\mathbf{K}_{1l}\right), & \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} &= \tfrac{1}{\tau_1^2}\mathbf{Z}_{1l}'\,\mathbf{r}_1, & &(\boldsymbol{\beta}_{1l}),
\end{aligned}
$$

where $\mathbf{r}$ is the current partial residual and $\mathbf{r}_1$ is the "partial residual" of the level-2 equation. More precisely, $\mathbf{r}_1 = \boldsymbol{\beta}_1 - \tilde{\boldsymbol{\eta}}_1$ and $\tilde{\boldsymbol{\eta}}_1$ is the predictor of the level-2 equation excluding the current effect of z_{1l}.

MCMC updates of the regression coefficients takes advantage of the following key features:

Reduced complexity in the second or third stage of the hierarchy: Updating the regression coefficients $\boldsymbol{\beta}_{1l}$, $l = 1, \ldots, q_1$, in the second (or third level) is done conditionally on the parameter vector $\boldsymbol{\beta}_1$. This facilitates updating the parameters for two reasons. First the number of "observations" in the level-2 equation is equal to the length of the vector $\boldsymbol{\beta}_1$ and therefore much less than the actual number of observations n. Second the full conditionals for $\boldsymbol{\beta}_{1l}$ are Gaussian regardless of the response distribution in the first level of the hierarchy.

Sparsity: Design matrices $\mathbf{Z}_j, \mathbf{Z}_{1l}$ and penalty matrices $\mathbf{K}_j, \mathbf{K}_{1l}$ and with it cross products $\mathbf{Z}_j'\mathbf{W}\mathbf{Z}_j, \mathbf{Z}_{1l}'\mathbf{Z}_{1l}$ and posterior precision matrices in (6) are often sparse. The sparsity can be exploited for highly efficient computation of cross products, Cholesky decompositions of posterior precision matrices and for fast solving of relevant linear equation systems.

Number of different observations smaller than sample size: In most cases the number m_j of different observations $z_{(1)}, \ldots, z_{(m_j)}$ in $\mathbf{Z}_j$ (or m_{1l} in $\mathbf{Z}_{1l}$ in the level-2 equation) is much smaller than the total number n of observations. The fact that $m_j \ll n$ may be utilized to considerably speed up computations of the cross products $\mathbf{Z}_j'\mathbf{W}\mathbf{Z}_j$, $\mathbf{Z}_{1l}'\mathbf{Z}_{1l}$, the vectors $\mathbf{Z}_j'\mathbf{W}\,\mathbf{r}$, $\mathbf{Z}_{1l}'\,\mathbf{r}_1$ and finally the updated vectors of function evaluations $\mathbf{f}_j = \mathbf{Z}_j\boldsymbol{\beta}_j$, $\mathbf{f}_{1l} = \mathbf{Z}_{1l}\boldsymbol{\beta}_{1l}$.

Sampling from the full conditionals (6) may still be a quite computer intensive procedure, particularly for many surface estimators. The prime example is a Gaussian random field (kriging) which is almost intractable in the standard parametrization. Therefore, we propose an alternative parametrization which simultaneously diagonalizes cross products and penalty matrices.

More specifically, for a particular function f let $\mathbf{Z}'\mathbf{Z} = \mathbf{R}\mathbf{R}'$ be the Cholesky decomposition of the cross product of the design matrix and $\mathbf{Q}\mathbf{S}\mathbf{Q}'$ be the singular value decomposition of $\mathbf{R}^{-1}\mathbf{K}\mathbf{R}^{-T}$.
Then the decomposition $\boldsymbol{\beta} = \mathbf{R}^{-T}\mathbf{Q}\tilde{\boldsymbol{\beta}}$ yields

$$\mathbf{Z}\boldsymbol{\beta} = \mathbf{Z}\mathbf{R}^{-T}\mathbf{Q}\tilde{\boldsymbol{\beta}} = \tilde{\mathbf{Z}}\tilde{\boldsymbol{\beta}},$$

where the transformed design matrix $\tilde{\mathbf{Z}}$ is defined by $\tilde{\mathbf{Z}} = \mathbf{Z}\mathbf{R}^{-T}\mathbf{Q}$. Now the cross product and penalty matrix of the transformed parameter vector $\tilde{\boldsymbol{\beta}}$ is diagonal resulting in computationally very favorable full conditionals. More details on our sampling schemes can be found in Lang et al. (2010).

4 Application on hedonic house price modeling

We analyze house price data belonging to three hierarchical levels of spatial units. Additionally to individual attributes, explanatory covariates with possibly nonlinear effects are available on two of these spatial resolutions. We use a dataset of 3917 owner-occupied single family homes in Austria and estimate a multilevel STAR model of the form:

$$\begin{aligned}
\text{level 1:} \quad & lnp && = f_{1,1}(area) + \cdots + f_{1,q_1}(age) + f_{spat_1}(s_1) + \boldsymbol{\varepsilon}_1 \\
\text{level 2:} \quad & f_{spat_1}(s_1) && = f_{2,1}(purchase\ power) + \cdots + f_{2,q_2}(level\ of\ education) \\
& && \quad + f_{spat_2}(s_2) + \boldsymbol{\varepsilon}_2 \\
\text{level 3:} \quad & f_{spat_2}(s_2) && = f_{3,1}(unemployment\ rate) + f_{spat_3}(s_3) + \boldsymbol{\varepsilon}_3 \\
\text{level 4:} \quad & f_{spat_3}(s_3) && = \beta_0 + \boldsymbol{\varepsilon}_4.
\end{aligned}$$

The top level equation is a STAR-model for logged home sales prices lnp with possibly nonlinear effects $f_{1,1}, \ldots, f_{1,q_1}$ of continuous structural house characteristics such as the floor space ($area$) or the age of the building (age). Spatial heterogeneity is modeled through the spatial random effect $f_{spat_1}(s_1)$ of municipalities s_1 which is further decomposed into a district and county level effect (spatial indexes s_2 and s_3). At levels 2 and 3 further possibly nonlinear effects $f_{2,1}, \ldots, f_{2,q_2}, f_{3,1}$ of locational characteristics are included. For technical reasons the global intercept β_0 is included on the lowest level.
Figures 1 and 2 exemplify some of the results. More details can be found in Brunauer et al. (2010).

References

Belitz, C., Lang, S. (2008). Simultaneous selection of variables and smoothing parameters in structured additive regression models. *Computational Statistics and Data Analysis*, **53**, 61-81.

Brunauer, W., Lang, S. and Umlauf, N. (2010). Modeling House Prices using Multilevel Structured Additive Regression. Technical report, University of Innsbruck.

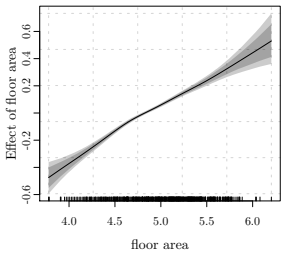


FIGURE 1. Nonlinear effects of floor area, age of the building and the purchase power index.

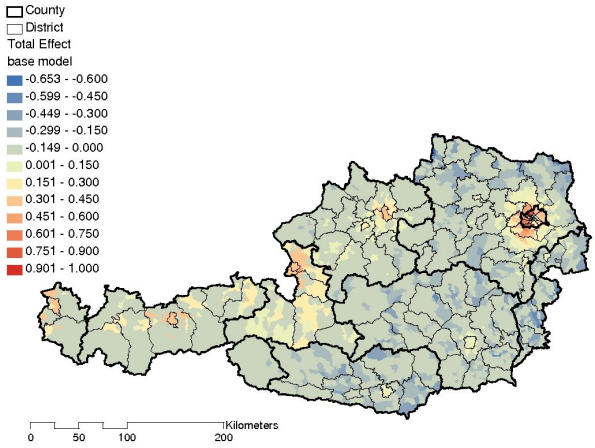


FIGURE 2. Total spatial effect $f_{spat_1}(s_1)$.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing using B-splines and penalized likelihood. *Statistical Science*, **11**, 89-121.

Fahrmeir, L., Kneib, T. and Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, **14**, 731-761.

Lang, S., Umlauf, N., Kneib, T., Hartgen, K. and Wechselberger, P. (2010): Multilevel Generalized Structured Additive Regression. Technical report, University of Innsbruck

Shadow Graphs for Contingency Tables

Joseph B. Lang[1]

[1] Department of Statistics and Actuarial Science, University of Iowa, USA

Abstract: Graphical methods for describing the goodness of fit of a model often play an important role in statistical analyses. Unfortunately, compared to models for interval- and ratio-level data, there are very few goodness-of-fit graphical methods for contingency table models. To address this problem, this paper introduces *shadow graphs*. Shadow graphs can be used to directly compare a wide variety of observed and expected features of contingency table data under a particular model. The broad applicability of shadow graphs is illustrated through two simple examples. Specifically, shadow graphs are used to assess the fit of several members of the class of multinomial-Poisson homogeneous models, which is a broad class that includes loglinear models and product-multinomial linear predictor models as special subclasses.

Keywords: Adjusted Residuals; Goodness of Fit; Multinomial-Poisson Homogeneous Models.

1 Introduction

1.1 Motivation

Graphical methods for describing the goodness (or lack) of fit of a model often play an important role in statistical analyses. Unfortunately, compared to models for interval- and ratio-level data, there are very few goodness-of-fit graphical methods for contingency table models.

To be clear, there are graphical displays of contingency table data, including the familiar bar and pie charts, and there are graphical displays of association among two or more categorical variables, including correspondence analysis plots and mosaic displays (cf. Friendly, 1994). These latter two can be viewed as graphical displays of the goodness of fit of the independence model, and extended versions of the mosaic display can be used to graphically display the goodness of fit of more general log-linear models. However, the author is not aware of any graphical method that displays the goodness of fit of more general contingency table models.

To address this problem, this paper introduces *shadow graphs*. Unlike mosaic displays, shadow graphs can be used to directly compare a wide variety of observed and expected features of contingency table data under a particular model. Like the mosaic display, the shadow graph includes residual information so that probabilistic assessments of lack of fit can be carried out.

TABLE 1. Siskel and Ebert Data

	*E*bert		
*S*iskel	con	mixed	pro
con	24	8	13
mixed	8	13	11
pro	10	9	64

TABLE 2. Gator Food Choice Data

	Primary Food Choice					Sample Estimate
Lake	Fish	Invert	Reptile	Bird	Other	of Dispersion*
Hancock	30	4	3	5	13	0.63
Oklawaha	18	19	7	1	3	0.68
Trafford	13	18	8	4	10	0.76
George	33	20	1	3	6	0.58

*Dispersion $D_i \equiv 1 - \sum_{j=1}^{5} \pi_{ij}^2$

The broad applicability of shadow graphs is illustrated through two simple examples. Specifically, shadow graphs are used to assess the fit of several members of the class of multinomial-Poisson homogeneous models, which is a broad class that includes loglinear models and product-multinomial linear predictor models as special subclasses.

1.2 Introduction to Examples

Example 1. Siskel and Ebert Rating Data.
Movie critics, Siskel and Ebert, rated 160 movies on the three point scale (con, mixed, pro). Table 1 displays the counts. To illustrate the use of shadow graphs we will assess the fit of three models: (1.1) Independence $[S \perp E]$, (1.2) Equal Means $[\mathcal{E}(S) = \mathcal{E}(E)]$, and (1.3) Zero Agreement $[Kappa = 0]$.

Example 2. Gator Food Choice.

Investigators recorded the primary food choice for alligators from four Florida lakes. The counts are displayed in Table 2. To illustrate the use of shadow graphs we will asses the fit of two models: (2.1) Independence $[Choice \perp Lake]$ and (2.2) Equal Dispersions $[D_i = \beta_0, \quad i = 1, 2, 3, 4]$

2 Model Specification and Fitting

Models (1.1)-(1.3) and (2.1)-(2.2) referred to in the examples of the previous section can all be viewed as multinomial-Poisson homogeneous (MPH) models as described in Lang (2004, 2005). An MPH model is a contingency table model that can be generically expressed as

$$y \leftarrow Y \sim MP(\nu, \pi|\ s, F), \quad \text{where} \quad h(\mu) = 0.$$

Here, y is the vector of table counts; ν is the vector of expected sample sizes; π is the vector of table probabilities; μ is the vector of table expected counts; s is a vector of strata identifiers; and F is the set of strata with a priori fixed sample sizes. The constraint function h is allowed to be any smooth vector-valued function that is homogeneous (see Lang 2004) with respect to the sampling plan. The MP notation tells us that the counts are modeled as a combination of multinomial and independent Poissons.
As examples, multinomial and Poisson loglinear models and the product-multinomial linear predictor models of Grizzle et al. (1969) are special case MPH models. In fact, any model of the form $F(\pi) = X\beta$, where F is a smooth link function, can be expressed as an MPH model, or more specifically a homogeneous linear predictor (HLP) model (Lang 2005).
The maximum likelihood fitting program `mph.fit`, which is written in `R` and available from the author, was used to fit the models of this paper. The shadow graphs were created, and the corresponding adjusted residuals of the generic form $(S(Y) - S(\hat{\mu}))/ase(S(Y) - S(\hat{\mu}))$ were computed, in `R`.

3 Shadow Graphs

Example 1. Siskel and Ebert. Figures 1 and 2 give shadow graphs that can be used to assess the goodness of fit of the independence model. Note that Figure 1 compares the observed (solid blue rectangles) and expected (gray shadows) joint distributions, whereas Figure 2 compares the observed and expected conditional distributions. It is clear from the shadow graphs that there are more counts in the diagonal cells than expected under the independence model. That is, the independence model is deemed untenable and a positive association is suggested.
Figure 3 gives two different shadow graphs for assessing the fit of the equal means model (1.2). The graph to the left compares the observed and expected joint distribution and the graph to the right compares the observed and expected means. The graphs do not indicate any statistical lack of fit of the model. In particular, the residuals are all small–none are bigger than 0.71 in absolute value.
Space considerations preclude the inclusion of shadow graphs for the third model, (1.3), considered in Example 1. Instead we move on to Example 2.

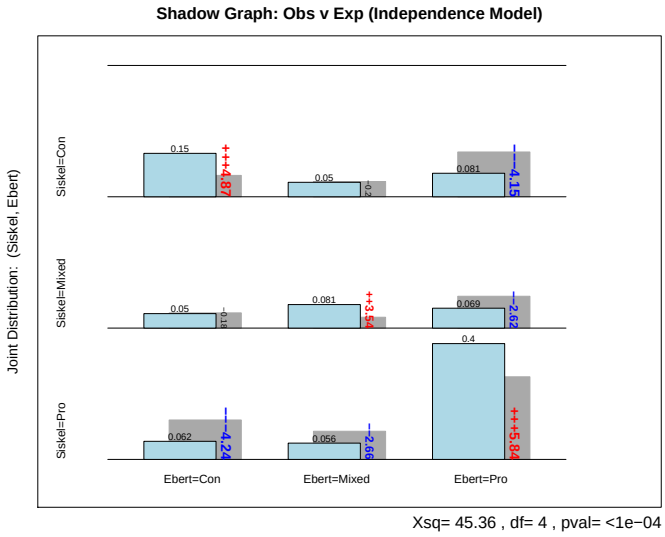


FIGURE 1. Shadow Graph: $S \perp E$ (Obs vs. Exp Joint Distns).

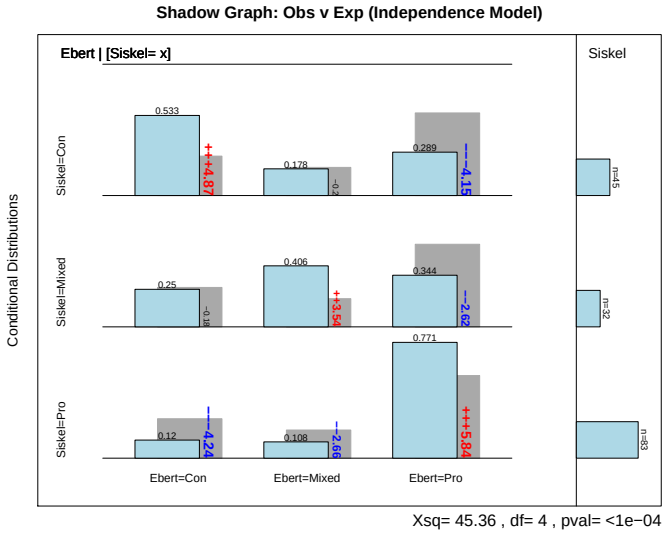


FIGURE 2. Shadow Graph: $S \perp E$ (Obs vs. Exp Conditional Distns).

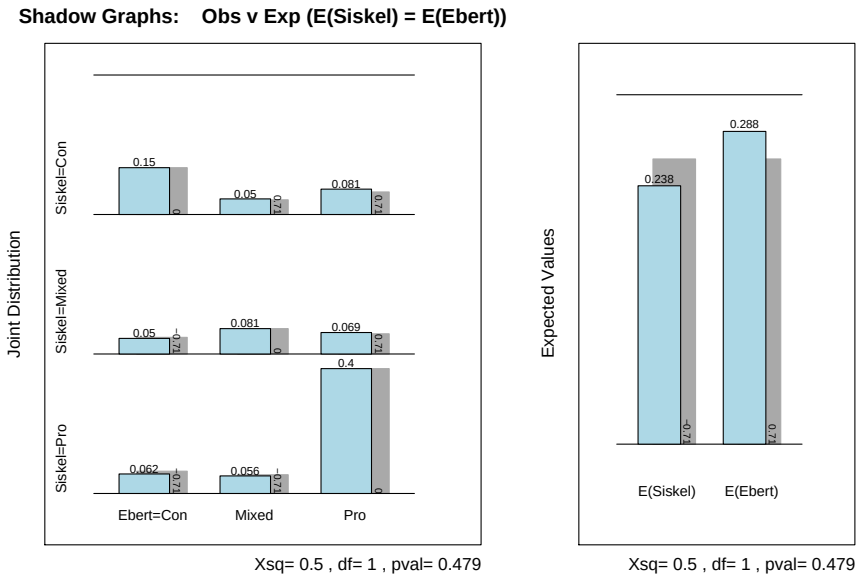


FIGURE 3. Shadow Graphs: $\mathcal{E}(S) = \mathcal{E}(E)$ (Obs vs. Exp Joint Distns, Obs vs. Exp Means).

Example 2. Gator Food Choice. Figures 4 and 5 give shadow graphs that can be used to assess the goodness of fit of the model of equal dispersions, which is a non loglinear model. Note that Figure 4 compares the observed (solid blue rectangles) and expected (gray shadows) dispersions, whereas Figure 5 compares the observed and expected conditional distributions. The shadow graphs show that the equal dispersion model is untenable. In particular, the dispersion in food choice is higher than expected for Lake Trafford and lower than expected for Lake George.

References

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, **58**, 190-200.

Grizzle, J.E., Starmer, C.F. and Koch, G.G. (1969). Analysis of categorical data by linear models. *Biometrics*, **25**, 489-504.

Lang, J.B. (2004). Multinomial-Poisson homogeneous models for contingency tables. *Annals of Statistics*, **32**, 340-383.

Lang, J.B. (2005). Homogeneous linear predictor models for contingency tables. *Journal of the American Statistical Association*, **100**, 121-134.

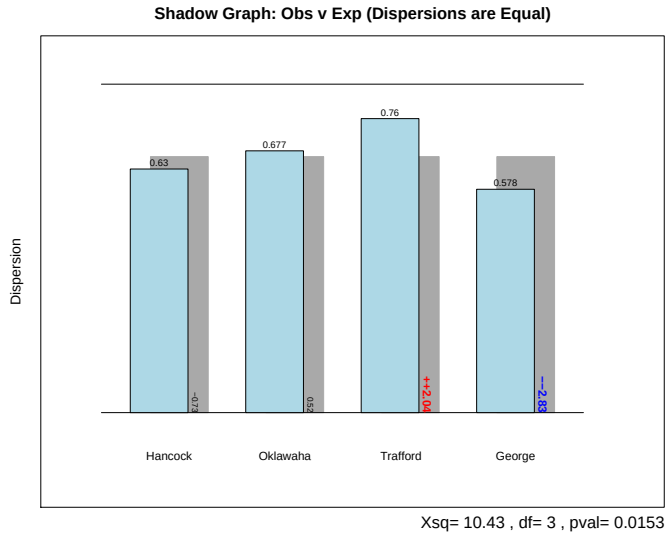


FIGURE 4. Shadow Graph: $D_i = \beta_0$ (Obs vs. Exp Dispersions).

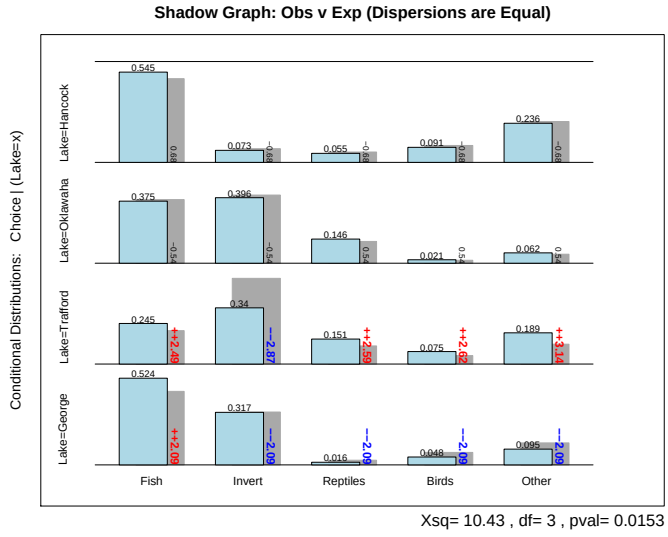


FIGURE 5. Shadow Graph: $D_i = \beta_0$ (Obs vs. Exp Conditional Distns).

How do the health risks from air pollution vary across communities in Scotland?

Duncan Lee [1]

[1] Department of Statistics, 15 University Gardens, University of Glasgow, Glasgow, G12 8QQ. email - d.lee@stats.gla.ac.uk

Abstract: The health risks from air pollution are likely to depend on the level of pollution and the vulnerability of the population. This paper investigates how these factors effect the pollution-health relationship across Scotland.

Keywords: Air pollution; Respiratory health; Varying-coefficient model.

1 Background

The health effects that result from long-term exposure to air pollution can be estimated using spatial ecological studies (see for example Lee *et al* (2009)), where the region of interest is split into contiguous small-areas. The size of these effects are likely to depend on a number of factors, including: (i) the level of pollution; (ii) the proportion of the population who are elderly; and (iii) the underlying health of the population. Therefore assuming there is a single health effect for the whole of Scotland is restrictive, and in this paper we investigate the characteristics that make populations more at risk. However, we do not model the health effects as a smooth function of location, because any spatial variation in risk is likely to be due to omitted risk factors, which themselves may not be spatially smooth.

2 Methods

We adopt a Bayesian hierarchical modelling approach that allows for spatial correlation, with inference based on Markov chain monte carlo (MCMC) simulation. The responses $\mathbf{y} = (y_1, \ldots, y_n)$ are the observed numbers of health events for each small-area over an extended period of time, meaning that a Poisson model is appropriate. In addition, we also have the expected number of health events in each area E_k, which are based on the demographics of the population and will be used as an offset term in the model. The observed admissions are regressed against air pollution concentrations $\mathbf{x} = (x_1, \ldots, x_n)$, and p covariate risk factors $Z = (\mathbf{z}_1^{\mathrm{T}}, \ldots, \mathbf{z}_n^{\mathrm{T}})$, the latter of which include measures of socio-economic deprivation, urbanicity and

demography. We allow the effects of air pollution to depend on an additional covariate $\mathbf{m} = (m_1, \ldots, m_n)$, termed an *effect modifier*, which yields a varying coefficient function $\beta(m_k)$ for the pollution-health relationship. To avoid specifying the functional form for $\beta(m_k)$, we represent it by a smooth function whose shape is estimated from the data. This function is modelled by a penalised B-spline (p-spline, Eilers and Marx (1996)), where an overly large set of basis functions is used, and smoothness is induced via a second order difference penalty. As we adopt a Bayesian approach to analysis the penalty is replaced by a second order random walk prior (Lang and Brezger (2004)), yielding the following model.

$$\begin{aligned} y_k &\sim \text{Poisson}(E_k R_k) \quad \text{for } k = 1, \ldots, n \\ \ln(R_k) &= \mathbf{z}_k^{\mathrm{T}} \alpha + x_k \beta(m_k) + \phi_k + \theta_k \\ \beta(m_k) &= \sum_{j=1}^{r} B_j(m_k) \gamma_j \\ \gamma_j &\sim \text{N}(2\gamma_{j-1} - \gamma_{j-2}, \sigma^2) \qquad \text{for } j = 3, \ldots, r \\ \phi_k | \phi_{-k} &\sim \text{CAR}(W, \tau^2) \\ \theta_k &\sim \text{N}(0, \nu^2) \end{aligned} \tag{1}$$

In this model $B_j(m_k)$ are the B-spline basis functions of the effect modifier, while the γ_j are the associated regression parameters. The smoothness of the p-spline is controlled by σ^2 (small values indicate greater smoothness), which is estimated within the MCMC inferential algorithm. The effects of unmeasured risk factors are modelled by a convolution of spatial (ϕ_k) and non-spatial (θ_k) random effects. The former is represented by a conditionally autoregressive (CAR) model, where the neighbourhood matrix, W, is based on areas sharing a common border. Model (1) is completed by the specification of Gaussian (α) and inverse-gamma (σ^2, τ^2 and ν^2) priors for the remaining parameters.

3 Air pollution and health in Scotland

3.1 Data description and study design

The data come from the Scottish Neighbourhood Statistics (SNS) database available from *http://www.sns.gov.uk/*, and the study region is the set of 1,235 Intermediate Geographies (IG) in Scotland (median population of 3,956). The health data are the numbers of respiratory admissions to hospital in 2005 in each IG, while the pollution data are modelled average PM_{10} concentrations between 2002 and 2004. The expected numbers of admissions were derived from data provided by the Information Services Division of the National Health Service. Numerous covariates are available

TABLE 1. Summary of the Scotland data.

Data	**Distribution**						
	0%	2.5%	25%	50%	75%	97.5%	100%
Admissions	10	24	40	54	73	121	181
PM_{10} (μgm^{-3})	8.4	9.1	11.6	13.1	14.1	16.2	19.3
Income deprived (%)	1.4	2.5	7.0	11.4	19.0	36.8	54.8
Pension population (%)	6.0	9.1	16.1	19.4	22.7	29.1	41.1
CHD SMR (%)	0.08	0.28	0.50	0.63	0.79	1.21	1.75

for this study, including measures of urbanicity, socio-economic deprivation, population health and population age structure, a summary of which is presented in Table 1.

3.2 Model building

Our model building process began by selecting the important covariates from the larger set that are available. Initially, all 39 available covariates were included in the model, and the importance of each covariate was assessed in turn using the deviance information criterion (DIC). As a result of this process a model with 24 covariates was selected, which corresponds to a reduction in the DIC from 10,356 to 10,341. These remaining covariates include ecological measures of socio-economic deprivation, urbanicity and population vulnerability, and include the percentage of the population who are income deprived, the proportion who are elderly, and the standardised morbidity ratio (SMR) for hospital admissions due to coronary heart disease (CHD).

The adequacy of this model was then checked by assessing its residuals, which did not exhibit any evidence of a spatial trend. However, they did exhibit substantial residual spatial correlation, which was assessed via a permutation test based on Moran's I statistic (a p-value less than 0.00001). This correlation was removed by adding a convolution of spatial and non-spatial random effects to the model, which resulted in a p-value of 0.9983 and reduced the DIC to 9,102.

3.3 Results

The results are based on 50,000 posterior samples, which were generated from 5 Markov chains that were burnt-in until convergence. Overall, the data show convincing evidence of a relationship between respiratory hospital admissions and PM_{10} concentrations, with an estimated relative risk

(for a one standard deviation ($1.8\mu gm^{-3}$) increase) of 1.067 with a 95% credible interval of (1.035, 1.101).

Figure 1 depicts how the relative risk of PM_{10} changes depending on (a) the percentage of the population that are pensioners; (b) the percentage of the population that are income deprived; (c) the SMR for coronary heart disease; and (d) the concentration of pollution. Panels (a), (b) and (c) present relative risks for a $1.8\mu gm^{-3}$ increase in PM_{10}, while panel (d) is the risk relative to the median PM_{10} level across Scotland.

Panel (d) of the figure shows that overall higher concentrations of PM_{10} result in larger estimated health risks, which is to be expected. Panel (a) shows that the health risks are worse for populations that are more elderly, which is again not surprising as they are more vulnerable than the general population. Finally, the estimated relative risks show quadratic relationships with income deprivation and coronary heart disease prevalence, with those in the middle of each distribution having the greatest risk. One explanation for this is that areas with low income deprivation and coronary heart disease are healthy, and thus are not affected by air pollution. Conversely, those areas at the opposite end of these distributions are very unhealthy, and as such are likely to be admitted to hospital for other causes than respiratory disease (thus reducing the observed relative risk).

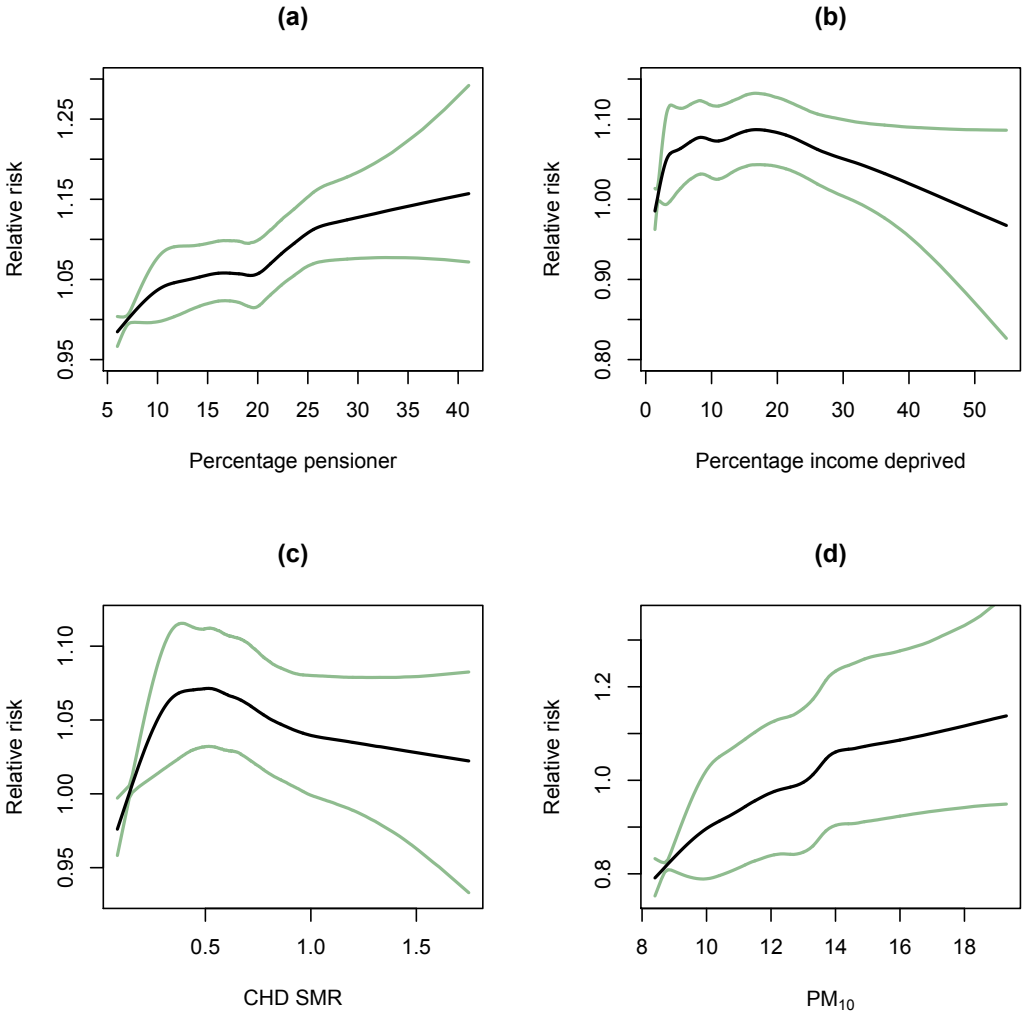


FIGURE 1. Relative risk functions for PM_{10} that depend on (a) proportion of pensioners; (b) levels of income deprivation; (c) the risk (SMR) of coronary heart disease; and (d) PM_{10} levels.

4 Future work

This short paper investigates how the health risks of air pollution vary across Scotland, focusing on the impact of pollution concentrations, de-

privation, ill health and demography. However, we have only considered 4 potential effect modifiers, and have included each in a separate model. Therefore future work will focus on how to select the important effect modifiers from a larger collection of potential risk factors, utilising Bayesian variable selection methods.

References

Eilers, P., and Marx, B. (1996). Flexible Smoothing with B-splines and Penalties *Statistical Science*, **11**, 89-121.

Lang, S., and Brezger, A. (2004). Bayesian p-splines *Journal of Computational and Graphical Statistics*, **13**, 183-212.

Lee, D., Ferguson, C., and Mitchell, R. (2009). Air pollution and health in Scotland: a multicity study *Biostatistics*, **10**, 409-423.

Spatial point pattern analysis: a multidimensional P-spline approach

Dae-Jin Lee[1], María Durbán[1]

[1] Department of Statistics, Universidad Carlos III de Madrid, SPAIN.
e-mail: dae-jin.lee@uc3m.es and mdurban@est-econ.uc3m.es

Abstract: In recent years, there have been an increasing interest in point patterns from many research fields (e.g. ecology, environmental studies or geographical epidemiology). Spatial point pattern data describes the spatial location events in a region or spatial domain. This work is motivated by the proliferation of georeferenced databases and the development of geographical information systems (GIS) software. They provide a variety of visualization and exploration tools. However, most of the models considered in a theoretical framework were developed in the kernel approach. We propose the use of penalized splines (P-splines) as a simple and effective tool for smoothing spatial point patterns. Our approach is essentially a bivariate P-spline Poisson regression, the spatial point patterns are processed and reduced to a $2d$ histogram. We illustrate the methodology with some examples.

Keywords: spatial point patterns; P-splines; multidimensional density estimation; GLAM.

1 Introduction

A point pattern consists in a set of locations, let's say $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, in a spatial domain or region, $\mathcal{D}$, where *events* of interest have been recorded. The analysis of spatial point patterns consider the *intensity function* as the average density points or the expected number of points per unit area. A simple solution to study the spatial pattern, is to count the number of events per unit area within a quadrat or window. A generalization of this idea is the use of non-parametric techniques to estimate the intensity function. Most of the techniques applied to this type of data have been developed from the kernel approach. This approach has many interesting mathematical properties and, in this context, can also consider non-stationary intensity functions. However, kernels suffer from some drawbacks as for example: *edge effects* at boundaries, that can lead to biased estimates.

In this paper, we consider the use of Penalized splines (Eilers and Marx, 1996) for the estimation of the intensity function. We consider the problem similar to estimate a density function, processing the point patterns and

reduced them to a $2d$ histogram. The estimation of the intensity function is achieved through a Poisson P-spline regression model, with counts of each bin as in Eilers (2006b). The attractive of this approach is the possibility of deal with a large amount of data and histograms of many (even thousands) bins, using the generalized linear array methods (GLAM) algorithms proposed by Currie et al. (2006) and Eilers et al. (2006a).

2 Two-dimensional P-splines for spatial point patterns

Let $\boldsymbol{s} = (\boldsymbol{s}_1, \boldsymbol{s}_2)$ be the pair of spatial locations where the events of study have been located. We can compute a $2d$ histogram with $n_1 \times n_2$ bins and form an array $\boldsymbol{Y}$ of Poisson counts in each bin, such that $\boldsymbol{y} = \text{vec}(\boldsymbol{Y})$, where $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are the midpoints of the bins in each spatial dimension. The expected values can be arranged in the array $\boldsymbol{M}$, and then, the mean is $\boldsymbol{\mu} = \text{vec}(\boldsymbol{M})$. The $2d$ P-spline Poisson model with log link and linear predictor is given by:

$$\boldsymbol{\eta} = \exp(\boldsymbol{\mu}) = \boldsymbol{B}\boldsymbol{\theta} \ , \tag{1}$$

where the regression basis $\boldsymbol{B}$ is the Kronecker product of the marginal B-spline basis calculated from $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, i.e. $\boldsymbol{B}_2 \otimes \boldsymbol{B}_1$, of dimensions $n_1 n_2 \times c_1 c_2$. We use cubic splines and equally spaced knots for the construction of the marginal basis.

The vector of coefficients, $\boldsymbol{\theta}$ can be arranged in a $c_1 \times c_2$ array $\boldsymbol{\Theta}$. Thus, (1) can be written as a GLAM:

$$\boldsymbol{\eta} = \exp(\boldsymbol{M}) = \boldsymbol{B}_1 \boldsymbol{\Theta} \boldsymbol{B}_2' \ , \tag{2}$$

We penalize the regression coefficients, $\boldsymbol{\theta}' \boldsymbol{P} \boldsymbol{\theta}$, with a penalty matrix $\boldsymbol{P}$, over rows and columns of $\boldsymbol{\Theta}$, given by:

$$\boldsymbol{P} = \lambda_1 \boldsymbol{I}_{c_2} \otimes \boldsymbol{D}_1' \boldsymbol{D}_1 + \lambda_2 \boldsymbol{D}_2' \boldsymbol{D}_2 \otimes \boldsymbol{I}_{c_1} \ , \tag{3}$$

where $\boldsymbol{D}_i$, $i = 1, 2$ is a second order difference matrix. It is worth mentioning that the penalty matrix in (3) allow for anisotropic processes, since it allows a different amount of smoothing (i.e. $\lambda_1 \neq \lambda_2$) in each spatial direction. This result is very important, because it is similar to consider non-separable covariance structures in a spatial process.

The estimation is done by maximization of the penalized Poisson log-likelihood:

$$\mathcal{L}_p = \mathcal{L} - \tfrac{1}{2} \boldsymbol{\theta}' \boldsymbol{P} \boldsymbol{\theta} \ , \tag{4}$$

where $\mathcal{L}$ is the ordinary log-likelihood. Maximizing (4) we obtain the system of equations:

$$\boldsymbol{B}'(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{P}\boldsymbol{\theta} \ , \tag{5}$$

which turn out to the penalized version of the scoring algorithm:

$$(\boldsymbol{B}'\tilde{\boldsymbol{W}}\boldsymbol{B}+\boldsymbol{P})\widehat{\boldsymbol{\theta}}=\boldsymbol{B}'\tilde{\boldsymbol{W}}\boldsymbol{B}\widetilde{\boldsymbol{\theta}}+\boldsymbol{B}'(\boldsymbol{y}-\widetilde{\boldsymbol{\mu}})\ , \tag{6}$$

where, $\tilde{\boldsymbol{\mu}}$, $\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{W}}$ denotes the current approximate solution, and $\widehat{\boldsymbol{\theta}}$ denotes the updated estimate of $\boldsymbol{\theta}$. The matrix $\boldsymbol{W}=\text{diag}(\boldsymbol{\mu})$. The algorithm (6), can be written as:

$$(\boldsymbol{B}'\tilde{\boldsymbol{W}}\boldsymbol{B}+\boldsymbol{P})\ \widehat{\boldsymbol{\theta}}=\boldsymbol{B}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}}\ , \tag{7}$$

where $\tilde{\boldsymbol{z}}=\tilde{\boldsymbol{\eta}}+\tilde{\boldsymbol{W}}^{-1}(\boldsymbol{y}-\tilde{\boldsymbol{\mu}})$ is known as the *working vector*. The smoothing parameters can be estimated using the Akaike's information criteria (AIC).

In many situations we might be interested in checking if the process has separable intensity function, that is, if $\mathbb{E}[\boldsymbol{Y}]$ is expressed as:

$$f(\boldsymbol{x}_1)+f(\boldsymbol{x}_2)\quad\text{or as}\quad f(x_1)+f(x_2)+f(x_1,x_2).$$

Recently, Lee and Durbán (2010) showed that it is possible to reparameterize the two-dimensional B-spline basis in such a way that the models above can be represented in terms of ANOVA-type decompositions. We propose the use of this representation, and an adaptation of the GLAM algorithms (Currie et al., 2006) in order to fit and compare these two models.

3 Examples

We consider the `lansing` data available in the R package `spatstat`. The data come from an investigation in Lansing Woods, Clinton County, Michigan USA. The data presents the locations of 2251 trees and their botanical classification (into hickories, maples, red oaks, white oaks, black oaks and miscellaneous trees). The original spatial locations have been rescaled to the unit square. A question of interest is to study the spatial pattern of the concentration of the trees species. We considered a $2d$ histogram of 30 bins for each spatial coordinate. Figure 1, shows in the left panel the $2d$ histogram of counts for the species of maples (with 514 observed trees). The smoothed the intensity function is shown in the right panel of Figure 1.

Concluding remarks

We presented a computationally efficient methodology for estimation of intensity functions in spatial point patterns using P-splines. The approach is similar to consider a bivariate density estimation of Poisson counts. GLAM methods allow for a fast and compact notation for implementation in standard software. The approach presented in this paper has very attractive features due to the possibility of extending the methods for spatio-temporal

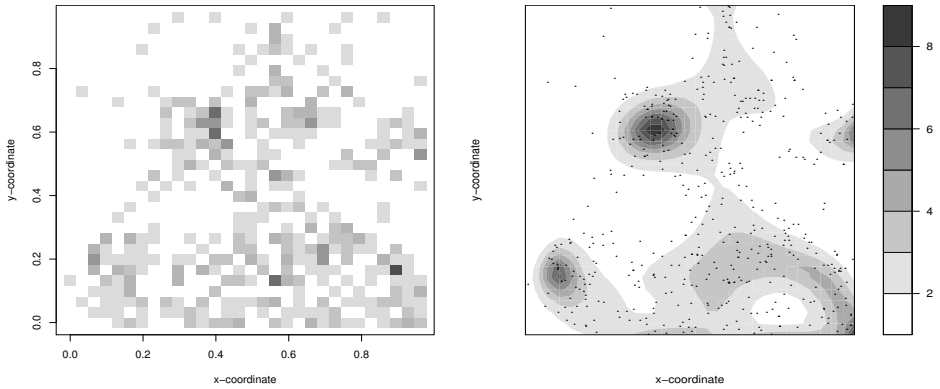


FIGURE 1. Left panel: counts of maples trees in a $2d$ histogram with 30-by-30 bins. Right panel: Smoothed intensity function for the spatial points pattern of maples, with B-spline basis of 8-by-8 knots.

point patterns, when we study events over time as for example: incidence of disease, births of species, ocurrences of fires or earthquakes. This might be an interesting alternative since traditional approaches in point pattern analysis ignore the incorporation of additional covariates or interactions due to model complexity.

Acknowledgements

The research of Dae-Jin Lee and María Durbán was supported by the Spanish Ministry of Science and Innovation (project MTM 2008-02901).

References

Currie, I. D., Durbán, M. and Eilers, P. H. C. (2006). Generalized linear array models with applications to multidimensional smoothing. *J. R. Statist. Soc. B*, **68**, 1-22.

Eilers, P. H. C., Currie, I. D., and Durbán, M. (2006a). Fast and compact smoothing on large multidimensional grids. *Computational Statistics & Data Analysis*, **50(1)**, 61–76.

Eilers, P. H. C., Marx, B. D. (2006b). Multidimensional density smoothing with P-splines. In proceeding of the *International Workshop on Statistical Modelling*, Galway (2006), pages: 151–158.

Lee, D.-J., and Durbán, M. (2010). P-spline ANOVA-Type interaction models for spatio-temporal smoothing. *to appear in Statistical Modelling.*

Copula based estimate of the likelihood ratio function for combining continuous biomarkers

Emilio Letón[1] , Elisa-María Molanes-López[2]

[1] Department of Artificial Intelligence, UNED, C/ Juan del Rosal 16, 28040 Madrid, Spain. E-mail: emilio.leton@dia.uned.es
[2] Department of Statistics, UC3M, Avda. de la Universidad 30, 28911 Leganés (Madrid), Spain. E-mail: elisamaria.molanes@uc3m.es

Abstract: In some biomedical studies, it is of special interest to develop methodologies that allow us to combine several biomarkers into a scalar-valued function that increases the classification accuracy of each biomarker alone. Based on the Neyman-Pearson lemma as a theoretical basis, we propose to combine multiple biomarkers using an estimate of a reparametrization of their likelihood ratio function via copula function estimates. Then, through a simulation study we show that the combined predictor outperforms the behaviour of each biomarker alone. Finally, a real dataset of cancer, well-known in the literature, is included to illustrate the techniques described in this paper.

Keywords: copulas; likelihood ratio function; relative distribution; ROC curve, Youden index.
Long abstract version: accepted for oral presentation.
Communicating author: Elisa-María Molanes-López.

1 Introduction

Continuous biomarkers are commonly used for classifying individuals into two groups of interest (for instance, the diseased and healthy populations). Without loss of generality, it is usually assumed that for a continuous biomarker, Y, and a given threshold c, $c \in \Re$, the individual with $Y > c$ is classified as diseased or positive and otherwise is classified as healthy or negative. For every fixed c, there are two indices of interest that are associated with the above-mentioned binary classification rule, the sensitivity or the proportion of 'true positive subjects', denoted by $q(c)$, and the specificity or the proportion of 'true negative subjects', denoted by $p(c)$. For a continuous biomarker, the accuracy is usually described graphically through the 'Receiver Operating Characteristic' (ROC) curve, which is obtained by plotting the pairs $(1 - p(c), q(c))$ as the threshold c is moved along the real line. In this setting, a commonly used global measure of classification performance is given by the area under the ROC curve which is denoted by AUC. However, in medical practice it is of main relevance

to know the threshold value of the biomarker to be used for classifying the individuals. There exist two main methods for identifying the optimal cut-off point: the northwest corner and the Youden index. We will focus on the latter method, recently studied by Schisterman and Perkins (2007) and Letón and Molanes-López (2009), among others.

In practice, it is unlikely that a single biomarker will detect the illness with both high sensitivity and specificity. Therefore, it is of main interest to know how to best combine multiple biomarkers into a score or scalar-valued function, that can be used for predicting the binary outcome of interest. In the literature, this problem has been tackled considering the linear combination of the available biomarkers that maximizes the empirical *AUC* (Pepe et al., 2006 and Ma and Huang, 2007). In this paper we relax this assumption using the likelihood ratio function.

From here on we will consider that there are p biomarkers available, i.e. a p-dimensional vector of biomarkers, $\mathbf{Y} = (Y_1, \ldots, Y_p)$, and we will use the notation $\mathbf{Y}_0$ and $\mathbf{Y}_1$ to refer to the p-dimensional biomarker in the healthy and diseased populations, respectively. Pepe (2003) establishes, with the help of the Neyman-Pearson lemma, the optimality of the likelihood ratio function defined by $LR(\mathbf{y}) = \mathbf{f}_1(\mathbf{y})/\mathbf{f}_0(\mathbf{y})$, where $\mathbf{y} = (y_1, \ldots, y_p) \in \Re^p$ and $\mathbf{f}_k$ refers to the multivariate density function of $\mathbf{Y}_k$, for $k = 0, 1$. This optimality refers to the fact that the ROC curve for any other score function is everywhere below the ROC curve for LR. Taking into account this result, we propose to estimate a reparametrization of LR into the p-dimensional unit cube and use it as a scalar-valued function for classification.

In Section 2, we first introduce the concept of relative distribution and relative density, and then, using Sklar's theorem we propose an estimate of a reparametrization of the likelihood ratio function of $\mathbf{Y}$, based on copula estimates, relative density estimates and relative distribution estimates. In Section 3, we study the performance of the combined biomarker, previously proposed, through a simulation study under different scenarios. Finally, Section 4 is devoted to illustrate the new methodology using a well-known real example.

2 New method

The relative distribution function of $Y_{1\ell}$ with respect to (wrt) $Y_{0\ell}$ is defined by

$$R_\ell(t_\ell) = \Pr(F_{0\ell}(Y_{1\ell}) \leq t_\ell) = \Pr(Y_{1\ell} \leq F_{0\ell}^{-1}(t_\ell)) = F_{1\ell}(F_{0\ell}^{-1}(t_\ell)),$$

where $t_\ell \in (0, 1)$, $F_{0\ell}$ denotes the cumulative distribution function (cdf) of $Y_{0\ell}$ and $F_{0\ell}^{-1}$ refers to the inverse of $F_{0\ell}$, for $\ell = 1, \ldots, p$. On the other hand, the relative density function of $Y_{1\ell}$ wrt $Y_{0\ell}$ is defined by

$$r_\ell(t_\ell) = \frac{\partial R_\ell(t_\ell)}{\partial t_\ell} = \frac{f_{1\ell}(F_{0\ell}^{-1}(t_\ell))}{f_{0\ell}(F_{0\ell}^{-1}(t_\ell))}.$$

Using Sklar's theorem (see Nelsen, 2006), the multivariate cdf of $\mathbf{Y_k}$, denoted by $\mathbf{F}_k$, and with univariate marginals given by $F_{k1}, \ldots, F_{kp}$, for $k = 0, 1$, can be rewritten by

$$\mathbf{F}_k(\mathbf{y}) = C_k(F_{k1}(y_1), \ldots, F_{kp}(y_p)), \tag{1}$$

where $\mathbf{y} = (y_1, \ldots, y_p) \in \Re^p$ and C_k denotes a copula function that describes the dependence structure between the marginals of $\mathbf{Y}_k$. Consequently, the multivariate density function $\mathbf{f}_k$, for $k = 0, 1$, can be rewritten as follows

$$\mathbf{f}_k(\mathbf{y}) = c_k(F_{k1}(y_1), \ldots, F_{kp}(y_p)) \prod_{\ell=1}^{p} f_{k\ell}(y_\ell), \tag{2}$$

where c_k refers to the density of the copula C_k given above. Using (2), it is easy to prove that

$$LR(\mathbf{y}) = \frac{c_1(R_1(t_1), \ldots, R_p(t_p))}{c_0(t_1, \ldots, t_p)} \prod_{\ell=1}^{p} r_\ell(t_\ell), \tag{3}$$

where t_ℓ is such that $F_{0\ell}(y_\ell) = t_\ell$, and $R_\ell(t_\ell)$ and $r_\ell(t_\ell)$ denote, respectively, the relative distribution and relative density of $Y_{1\ell}$ wrt $Y_{0\ell}$, for $\ell = 1, \ldots, p$. Note that (3) can be seen as a reparametrization of LR in the p-dimensional unit cube. Therefore, in order to estimate $LR(\mathbf{y})$, we propose to separately estimate every univariate relative distribution and relative density involved in (3) using kernel estimators (see Molanes-López and Cao, 2008, among others), and then to fit the copula functions C_0 and C_1 appearing in (1) using a family of parametric copulas.

3 Simulation study

A study of the performance of the combined biomarker obtained with the methodology proposed in Section 2 is carried out here when several biomarkers are available. We consider different scenarios for generating values of the multidimensional biomarker $\mathbf{Y}$, using copula functions to model the dependence structure existing between the marginals of $\mathbf{Y}$. Regarding the marginals, we include a variety of parametric models similar to those considered by other authors in the literature. However, for the sake of brevity, we only present here the results regarding the scenarios where $n_0, n_1 = 100$, $p = 2$, the dependence structure is modeled by a bivariate Clayton copula (see Clayton, 1978), $C(\theta)$, with $\theta = 0.2941, 1, 2.8820, 4.9654$, and the marginals are normal distributed with $\sigma^2_{Y_{k\ell}} = 0.25$, for $\ell = 1, 2$, $k = 0, 1$, $\mu_{Y_{01}} = \mu_{Y_{02}} = 2.5$, and $\mu_{Y_{11}}$, $\mu_{Y_{12}}$ are accordingly selected to obtain $AUC_\ell = 0.60, 0.75, 0.90$, for $\ell = 1, 2$,

θ	AUC_2	$\widehat{AUC}_c$	$\widehat{AUC}_{LR}$	$\widehat{J}_1$	$\widehat{J}_2$	$\widehat{J}_c$	$\widehat{J}_{LR}$
0.2941	0.60	0.659	0.633	0.167	0.169	0.253	0.213
	0.75	0.767	0.757	0.170	0.388	0.413	0.395
	0.90	0.903	0.900	0.170	0.648	0.659	0.649
1	0.60	0.653	0.624	0.169	0.169	0.244	0.197
	0.75	0.772	0.759	0.170	0.388	0.424	0.389
	0.90	0.915	0.918	0.169	0.648	0.707	0.692
2.8820	0.60	0.643	0.622	0.169	0.167	0.227	0.191
	0.75	0.823	0.817	0.170	0.388	0.533	0.515
	0.90	0.943	0.961	0.168	0.646	0.829	0.833
4.9654	0.60	0.639	0.626	0.169	0.170	0.220	0.197
	0.75	0.878	0.875	0.168	0.385	0.643	0.635
	0.90	0.949	0.979	0.170	0.647	0.878	0.902

TABLE 1. *Average AUC and J over 1000 trials with* $AUC_1 = 0.60$.

where

$$AUC_\ell = \Phi\left(\frac{\mu_{Y_{1\ell}} - \mu_{Y_{0\ell}}}{\sqrt{\sigma^2_{Y_{1\ell}} + \sigma^2_{Y_{0\ell}}}}\right),$$

(see Pepe, 2003) and Φ denotes the cdf of the standard normal.

θ	AUC_2	$\widehat{AUC}_c$	$\widehat{AUC}_{LR}$	$\widehat{J}_1$	$\widehat{J}_2$	$\widehat{J}_c$	$\widehat{J}_{LR}$
0.2941	0.60	0.769	0.756	0.389	0.171	0.415	0.394
	0.75	0.820	0.809	0.386	0.387	0.501	0.483
	0.90	0.914	0.910	0.387	0.647	0.683	0.675
1	0.60	0.772	0.760	0.388	0.170	0.424	0.390
	0.75	0.802	0.792	0.386	0.388	0.467	0.451
	0.90	0.908	0.906	0.385	0.648	0.668	0.653
2.8820	0.60	0.821	0.816	0.388	0.170	0.528	0.511
	0.75	0.799	0.790	0.384	0.385	0.466	0.441
	0.90	0.916	0.931	0.385	0.650	0.739	0.729
4.9654	0.60	0.879	0.875	0.388	0.169	0.643	0.636
	0.75	0.799	0.795	0.387	0.387	0.479	0.453
	0.90	0.914	0.952	0.386	0.648	0.793	0.814

TABLE 2. *Average AUC and J over 1000 trials with* $AUC_1 = 0.75$.

It is interesting to note here that although the dependence structure has been modeled through a Clayton copula, we fit the data to different parametric copulas and choose the one that achieves the maximum likelihood value. The family of copulas that we have considered is composed of the

Gaussian copula, the Clayton copula, the rotated Clayton copula, the Plackett copula, the Gumbel copula and the rotated Gumbel copula.
In Tables 1 and 2 we collect the mean AUC values for the composite biomarker ($\widehat{AUC}_c$) and for the theoretical LR biomarker ($\widehat{AUC}_{LR}$), and also the mean Youden index, J, for the two univariate biomarkes, Y_1 and Y_2, the composite one, and the LR biomarker ($\widehat{J}_1$, $\widehat{J}_2$, $\widehat{J}_c$ and $\widehat{J}_{LR}$, respectively). Note that the larger the value of AUC and J, the better the biomarker performs. Consequently, from the results obtained in this simulation study, the outperformance of the combined biomarker proposed in the previous section is confirmed. These simulations have been carried out in MATLAB using 1000 trials per scenario.

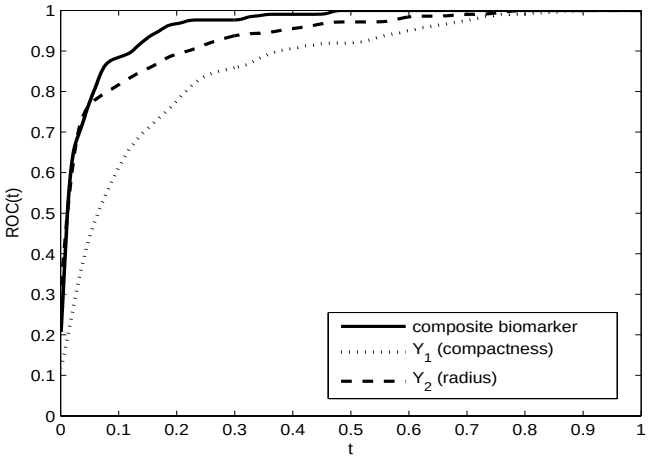


FIGURE 1. *ROC curve of Y_1 (dotted line), Y_2 (dashed line) and composite biomarker (solid line).*

4 Example

In this section, we illustrate the use of the new methodology to combine multiple biomarkers into a composite one, using the Wisconsin Diagnostic Breast Cancer dataset, publicly available in the repository at http://archive.ics.uci.edu/ml/machine-learning-databases/.
This dataset, that has also been analyzed in Jain and Abraham (2003), among others, consists of 569 individuals, 212 out of them suffering from malignant breast cancer and 357 without the disease. For every individual there is available information on several continuous biomarkers, however we will only consider two of them, the average compactness (Y_1) and the average radius (Y_2).

After applying the methodology proposed in Section 2, we are able to combine these two biomarkers into a composite one that outperforms the behaviour of each alone (see Figure 1), achieving high specificity and sensitivity simultaneously, which is desirable in screening studies (see Pepe, Etzioni et al., 2001).

References

Clayton, D.G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, **65**, 141151.

Jain, R. and Abraham, A. (2003). A comparative study of fuzzy classification methods on breast cancer data. In: *Proceedings of the 7th International Work Conference on Artificial and Natural Neural Networks*. 512-519, Menorca, Spain.

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

Ma, S. and Huang, J. (2007). Combining multiple markers for classification using ROC. *Biometrics*, **63**, 751-757.

Molanes-López, E.M. and Cao, R. (2008). Plug-in bandwidth selector for the kernel relative density estimator. *Annals of the Institute of Statistical Mathematics*, **60**, 273-300.

Nelsen, R.B. (2006). *An introduction to copulas*. Second Edition. New York: Springer Verlag.

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

Pepe, M.S., Cai, T. and Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*, **62**, 221-229.

Pepe, M.S., Etzioni, R., et al. (2001). Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute*, **93**, 1054-1061.

Schisterman, E.F. and Perkins, N.J. (2007). Confidence intervals for the Youden index and corresponding optimal cut-point. *Communications in Statistics - Simulation and Computation*, **36**, 549-563.

Parametric survival models in the presence of informative censoring

Francesca Little[1], Thalia Petousis[1]

[1] Department of Statistical Sciences, University of Cape Town, Private Bag, Rondebosch 7701, South Africa. *e-mail*:francesca.little@uct.ac.za

Abstract: This paper describes the use of parametric survival models and inverse proportional censoring weights to estimate the duration of gametocytemia in patients with moderate malaria, in the presence of informative right censoring.

Keywords: Survival models; Inverse proportional censoring weights; informative censoring; gametocytemia.

1 Introduction

Gametocytes are the sexual form of the malaria parasite and the main agents of transmission. Malaria is diagnosed based on the presence of asexual parasites and most standard malaria drugs act on these with the aim of eliminating them. However, of vital importance to contain malaria and prevent its spread is the prevention and/or curtailment of gametocytemia. Studies are designed to measure asexual parasites. Gametocyte emergence usually lags parasite emergence and hence these studies have incorrect intervals for measurement of gametocytemia. In addition patients are withdrawn if parasites do not clear or if they re-emerge. In this way the failure to clear asexual parasites is a competing event to the measuring of gametocytes, with the result that we do not see the gametocyte distributions or we do not see the complete gametocyte distributions for patients who are resistant to the drug and hence are withdrawn due to treatment failure. Of interest is to determine the time to gametocytemia and the duration of gametocytemia. This paper looks at models for time to event data to provide estimates of duration of gametocytemia and the incorporation of inverse probabilty of censoring weights to adjust for possible biases due to missing data as a result of the inherent censoring process.

The data comes from a series of randomized clinical trials conducted between 2002 and 2004 in Mpumulanga (South Africa) and in five different centers in Mozambique as part of the SEACAT evaluation of the phased introduction of combination anti-malarial therapy that form part of the Lubombo Spatial Development Initiative malaria control programme

(Sharp *et al.* ,2007). Those diagnosed with pure uncomplicated acute P. falciparum malaria parasitaemia and who lived close enough to the study site for reliable follow up, were deemed suitable for inclusion. Enrolled subjects were seen on day 0 and then asked to return to the clinic on days 1, 2, 3, 7, 14, 21, 28 and 42 for assessments relating to clinical and parasitological end points.

2 Statistical Models

We overcome the bias due to censoring by calculating weights inversely proportional to the probability of not being censored so as to equalize the contribution of the number of subjects in the different strata at each visit. The calculation of these weights follow those described in Fewell *et al.*, (2004) as follows:

$$W(t) = \Pi_{k=0}^{t} p\{C(k) = 0 | \overline{C}(k-1) = 0, V, T > k\}$$

where $C(k)$ is a binary variable that equals 1 if censored on day k, and zero otherwise, $\overline{C}(k-1)$ is the event history of censoring, V are the baseline covariates, and T is the time variable. In practice, we estimate $p\{C(k)|V\}$ at each visit using a logistic regression model and then successively accumulate the complement of these probabilities over visits.

We estimate the probability of carrying gametocytes using a logistic model and estimate duration of gametocytemia by fitting a weibull model for time to gametocyte clearance in the accelerated failure time metric because of its formulation of the response in terms of time (Collett, 2003). The Survivor function is thus estimated using

$$S_i(t) = exp(-\lambda_i t^{1/\sigma})$$

where $\lambda_i = exp\{-(\mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \ldots + \alpha_p x_{pi})/\sigma\}$ and x_{ji} refers to the predictor variables.

Both the logistic model for gametocyte prevalence and the weibull regression model for gametocyte duration are weighted using the weights discussed above. Estimates for mean duration of gametocytemia among patients for whom we did observe gametocytemia are then obtained by estimating the area under the survival curve,

$$s(t) = \int_0^\infty S(t)dt.$$

Finally we combine the estimates of gametocyte prevalence and gametocyte duration among those who did carry gametocytes to obtain an estimate of gametocyte duration in the population as follows:

$$duration = p\{prevalence\} \times s(t)$$

We obtain confidence intervals for these estimates of gametocyte duration through bootstrapping.

3 Results

Table 1 illustrates the imbalance with respect to the distribution of subjects in the parasitic outcome categories over time and shows how at the later visits, we have fewer parasitic failures and those lost to follow up. To a lesser extent, a similar imbalance was observed with respect to the distribution of subjects in the different mutation categories over time.

TABLE 1. Observed subject distribution by outcome and time.

Day	Success	Failure	LTFU	Total
0	569	80	85	734
	77.52	10.90	11.58	100.00
3	564	70	57	691
	81.62	10.13	8.25	100.00
7	559	65	54	678
	82.45	9.59	7.96	100.00
14	550	49	35	634
	86.75	7.73	5.52	100.00
21	530	41	27	598
	88.63	6.86	4.52	100.00
28	534	26	17	577
	92.55	4.51	2.95	100.00
42	555	16	7	578
	96.02	2.77	1.21	100.00

Inverse proportional censoring weights are derived using logistic regression models at each time point that relates the logit probability of censoring to mutation and outcome information. Stratified probabilities of censoring at each time point are generated from these models,

$$p_i = exp(\beta_0 + \beta_1 out_i + \beta_2 mutcat_i)/(1 + exp(\beta_0 + \beta_1 out_i + \beta_2 mutcat_i))$$

for timepoint j. The complement of these probabilities are then accumulated over time to calculate probabilities of still being in the study for each patient at each successive time point, and the weights are calculated as the inverse of these cumulative probabilities, $w_{ij} = \frac{1}{\Pi_{k=1}^{j}(1-p_{ij})}$.

Figure 1 illustrates the time-varying weights for the different outcome and mutation strata. Weights stay relatively small until day 28 and the LTFU resistant strata have the larger weights.

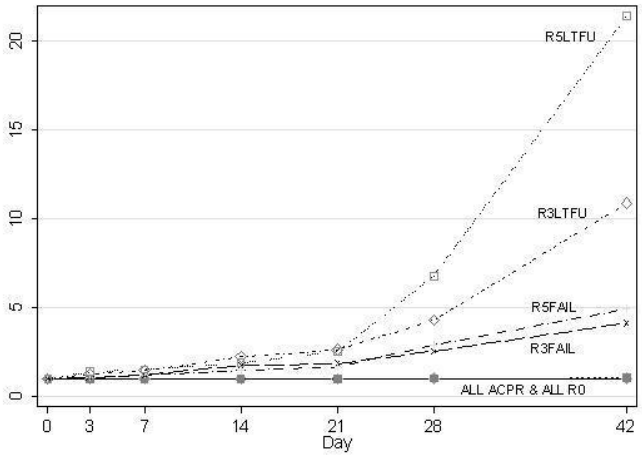


FIGURE 1. Censoring weights by outcome and mutation strata.

The counterfactual populations at each time point are then created by weighting analyses by these probabilities, resulting in an equal distribution of subjects over outcome groups at each time point, Table 2.

TABLE 2. Weighted Subject distribution by outcome and time.

Day	Success	Failure	LTFU	Total
0	569	80	85	734
	77.52	10.90	11.58	100.00
3	536	74.80	71.40	682.20
	78.57	10.96	10.47	100.00
7	531	76.76	76.03	683.79
	77.66	11.23	11.12	100.00
14	524.94	74.40	74.76	674.09
	77.87	11.04	11.09	100.00
21	507.75	68.78	70.16	646.69
	78.51	10.64	10.85	100.00
28	513.66	71.87	80.32	665.85
	77.14	10.79	12.06	100.00
42	541.99	73.94	86.65	702.57
	77.14	10.52	12.33	100.00

Weibull accelerated failure time models that contain outcome, mutation, treatment, site and asexual parasite density variables are fitted with and without these time-dependent weights and estimates of duration of gametocytemia for patients for whom we did observe gametocytes are obtained by integrating the survival functions, as described above. A logistic regression model is fitted to estimate the probabilities of carrying gametocytes using the same set of predictor variables. A weighted version of the logistic model is fitted using the estimated weight at the time of censoring for each patient. The estimated durations from the weibull models are then adjusted using these predicted probabilities to account for the patients for whom we never observed gametocytemia.

These models result in estimates of gametocyte duration for the overall population and by treatment, parasitic outcome and resistant mutation category. A comparison of the estimates of duration from successive models illustrates the relative success of the use of weights and prevalence probabilities in removing the biases due to censoring that results from the design of the trials yielding our data. Table 3 illustrates this process for the outcome and mutation stratifications. We note the decrease in duration after adjustment by the probability of gametocyte prevalence. The effect of the weighted analysis is most noticeable for the LTFU stratum, which is of course the stratum that had the largest weights.

TABLE 3. Estimated Durations

Stratum	Duration	Weighted Duration	Adjusted Duration	Adjusted weighted Duration
Outcome				
ACPR	21.95	21.88	12.94	12.63
	(21.46; 22.45)	(21.37; 22.38)	(12.57; 13.32)	(12.29; 12.97)
FAILURES	25.69	25.79	17.92	17.53
	(24.12; 27.25)	(24.37; 27.22)	(16.55; 19.28)	(16.34; 18.71)
LTFU	25.44	29.23	11.66	18.83
	(21.68; 29.20)	(24.78; 33.68)	(9.21; 14.11)	(15.08; 22.58)
Mutation				
0	27.46	27.53	14.26	14.12
	(27.27; 27.65)	(27.34; 27.73)	(14.10; 14.42)	(13.95; 14.28)
1-3	20.48	20.92	12.08	12.35
	(19.87; 21.09)	(20.24; 21.61)	(11.59; 12.57)	(11.80; 12.89)
4-5	23.24	22.77	15.59	15.23
	(22.21; 24.26)	(21.75; 23.80)	(14.69; 16.49)	(14.36; 16.10)

4 Discussion

We have shown how to overcome two biases inherent in the estimation of the duration of gametocytemia. Firstly we adjusted the estimated durations using the prevalence of gametocytemia to take into account the subset of the population who do not develop gametocytemia. In doing so, we have developed an approach not unlike the hurdle models (Min & Agresti, 2005) used for zero-inflated count data. Secondly we used inverse proportional censoring weights to correct for premature termination of distributions as a result of withdrawal of patients who fail treatment or are lost to follow up. The weighted analysis did not prove as useful as initially expected in correcting for possible biases due to early censoring.We noted that those who are censored earlier have higher observed gametocyte densities at the earlier time points. If this trend continues beyond the time of censoring there is a possibility that the profiles at the later timepoints for those patients who were censored earlier on may not be similar to the observed profiles at these points for the patients who were not censored. If this is the case, then weights that effectively multiply observed profiles of the completers may not be the optimal method of imputation to use.

Acknowledgments: The authors would like to thank Professor Karen Barnes, Division of Pharmacology, UCT and principal investigator of the SEACAT evalution, for her valuable clinical input and for allowing us to use her data.

References

Collett D. (2003). *Modelling Survival Data in Medical Research.* Second edition. Chapman& Hall/CRC.

Fewell Z., Herman M.A., and Wolfe F, Tilling K, Choi H, Sterne J (2004). Controlling for time-dependent confounding using marginal structural models. *The Stata Journal*, **4**, 402-420.

Min Y., and Agresti A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, **5**, 1-19.

Sharp BL, Kleinschmidt I., Streat E., Maharaj R., Barnes K.I., Durrheim D.N., Ridl F.C., Morris N., Seocharan I., Kunene S., La Granga J.J.P., Mthembu J.D., Maartens F., Martin C.L., Barreto A. (2007).

Seven years of regional malaria control collaboration Mozambique, South Africa and Swaziland. *Am J Trop Med Hyg* **76(1)**, 42-72.

Nonlinear transformations models: Application to mortality of calves.

L. López-Segovia[1], A. Espinal[2], G. Gómez[3]

[1] Ciencias Básicas, UJAT
[2] Servei d'Estadística, UAB
[3] Statistics Dept., UPC

Abstract: We present an application of the proportional hazard-proportional hazard cure models to evaluate the environmental and genetic factors that affect both mortality and survival up to weaning of beef calves. We use the ntlm package (software R) to fit nonlinear semiparametric transformation models, which include the proportional hazard cure models as a special case. Results indicate that genetic factors such as calving month and calving difficulty are associated with the long-term survivorship of the calves.

Keywords: Proportional hazard cure (PHC) model; Extended hazard (EH) models; Proportional hazard-Proportional hazard cure (PH-PHC) model; Nonlinear transformations (NLT) models.

1 Introduction

Mortality of calves from birth to weaning (approximately at 180 days) reduces farm's income, and significantly increases cattle production costs (see Goyache et al. (2003) for a review). Thus, it is important to take into account the survival pattern of calves into the overall breeding. Tarrés et al. (2005) used a standard survival analysis to study how genetic and environmental factors influence mortality up to weaning. However, and due to the high proportion of censoring in the data, one could think of the presence of a mixture of two subpopulations of calves: those susceptible to die before weaning and those who don't. A binary mixture model, also known as cure model, (e.g. Farewell, 1982) which takes into account a fraction of *cure* individuals, could be appropriate in this situation. In this paper we show an application of the extended hazard (EHM) models proposed by Tsodikov (2002) which are developed to combine both *long-term* and *short-term* effects. EHM models include as a particular case the proportional hazard cure models. We fit a proportional hazard - proportional hazard cure (PH-PHC) model to fit both genetic and environmental factors and discriminate between mortality of calves effects (*short-term* effects) and survival or cure effects (*long-term* effects).

2 Nonlinear transformation model

Let T be a non-negative random variable denoting the failure time of interest, with improper survival function $S_p(t|z)$ and bounded cumulative hazard function $H_p(t|z)$ such that $\pi(z) = S_p(\infty|z) > 0$ and $\theta(z) = H_p(\infty|z) < \infty$ and where z represents a vector of covariates. A model that takes into account the cure fraction $\pi(z)$ can be formulated into two ways, (i) as a mixture cure model (Farewell (1982)) given by

$$S_p(t|z) = \pi(z) + [1 - \pi(z)]S(t|z), \tag{1}$$

where $S(t|z)$ is the survival function for the time to failure conditional upon ultimate failure with $S(\infty|z) = 0$; (ii) by specifying a bounded cumulative hazard function $H_p(t|z)$ of the population (Tsodikov (2002)) and representing the survival function of T as

$$S_p(t|z) = \exp\{-\theta(z)F(t|z)\}, \tag{2}$$

where $F(t|z) = \frac{H_p(t|z)}{H_p(\infty|z)}$. In terms of the estimation of the cure fraction $\pi(z)$, the two representations ((1) and (2)) are equivalent within a nonparametric framework. Model (1) does not have the proportional hazard property, however when F does not depend on z, model (2) has the proportional hazard property and is refered as the proportional hazard cure model (PHC) (Tsodikov (2003a)).

The standardized cumulative hazard function $F(t|z)$, itself a distribution function, might depend on the covariate vector z. Thus, its corresponding survival function, $1-F(t|z)$, can be specified as a parametric transformation of the baseline survival function S_0 (representing a reference group of individuals) in terms of a second predictor $\eta(z)$ (Tsodikov (2003a), (2003b)). In particular, Lehmann alternatives for $1 - F(t|z)$ can be assumed, that is, $1 - F(t|z) = S_0^{\eta(z)}(t)$, yielding a PH model for $1 - F(t|z)$. The combined PH-PHC model is given by

$$S_p(t|z) = \exp\{-\theta(z)[1 - S_0^{\eta(z)}(t)]\}, \tag{3}$$

and allows separate modeling of the combined effects (*long-term* and *short-term* effects) of the covariate vector z, which may not necessarily be the same set for each predictor. This extension encompasses the PHC model when there are not short-term predictors, thats is, when $\eta(z) = 1$ and the PH model when there are not long-term predictors, that is, when $\theta(z) = 1$. We are assuming $\eta(z) = \exp(\beta_\eta z)$ and $\theta(z) = \exp(\beta_\theta z + \beta_c)$, hence β_η and β_θ are the regression coefficients for short-term effects and for long-term effects, respectively, and β_c is an additional regression parameter for the reference category of the cure fraction.

Inference procedures for regression coefficients β_η, β_θ and β_c are based on the generalized log-likelihood for a non linear transformation model. The

R-package `nltm` includes the PH-PHC model, among others, and uses restricted Nonparametric Maximum Likelihood Estimation procedure (Tsodikov (2002), (2003b)) to get parameter estimates.

3 Application: mortality and survival up to weaning of beef calves

Data includes characteristics of 2504 calves recorded between 1994 and 2002. The birth-weaning is the period that begins at birth and lasts during the first 180 days of life. Survival time is defined as the difference between the date of death and the date of birth. For those alive calves at the end of the observed period the survival time is right-censored. One of the main characteristics of the data is the high percentage of censoring: among the 2504 calves, only 68 were uncensored observations (2.7% dead calves). The dataset includes variables on the time of calving such as: cow's length of productive life at calving; calf birth weight and gender; month and year of birth, difficulties at calving as well as the herd to which the cow belongs. In the analysis presented in this paper we have excluded a total of 427 records: 168 records, corresponding to births in 1994 and 2002, were excluded because an irregular distribution of calves in the herds, and 259 records were excluded due to an insufficient follow-up (censored at $t = 1$ day). We analyze the remainder 2077 calves born between 1995 and 2001 from three different herds, with a total of 68 uncensored observations (3.27% dead calves). Mortality of calves from birth to weaning is characterized by a considerable fraction of cure or long-term survivors and this is shown in the heavy percentage of censoring in each herd: we encounter 94.77% alive calves in herd 1, 97.57% in herd 3 and 97.51% in herd 7. This is an empirical indication of the presence of a proportion of calves which are not susceptible to die before 180 days, and the use of a standard survival analysis can not be the most appropriate (Sy and Taylor, 2000).

4 Results

A sample of 2077 calves in three different herds has been analyzed. Descriptive statistics are summarized in Table 1. Included covariates were the length of productive live of the cow, say `lpl`, dicotomized into groups < 1300 *days* and > 1300 *days*, month of birth, say `month`, dicotomized into groups *September to Febrary* and *March to August*), gender (female, male) and the type of difficulties at calving, say `difficulty`, categorized into *without assistance* (reference group), *slightly assisted by the farmer* and *strongly assisted by the farmer or the veterinary practitioner*.
Due to heterogeneity among the three herds, separate PH-PHC models (as in (3)) were fitted for each herd. Table 2 displays the results for those models. Concerning long-term (cure) effects we find that calving month

TABLE 1. Characteristics of Calves

Factors	herd1 (%)	herd3 (%)	herd7 (%)
`Lpl`			
$< 1300d$	360 (58.72)	263 (63.52)	498 (47.42)
$> 1300d$	253 (41.27)	151 (36.47)	552 (52.57)
`Month`			
sep-feb	204 (33.27)	319 (77.05)	650 (61.90)
mar-aug	409 (66.72)	95 (22.94)	400 (38.09)
`Gender`			
female	308 (50.24)	216 (52.17)	544 (51.80)
male	305 (49.75)	198 (47.82)	506 (48.19)
`Difficulty`			
without assistance	299 (48.77)	379 (91.54)	916 (87.23)
slightly assisted	38 (6.19)	25 (6.03)	50 (4.76)
strongly assisted	3 (0.48)	9 (2.17)	83 (7.90)
missing	273 (44.53)	1 (0.24)	1 (0.09)
total	613 (29.51)	414 (19.93)	1050 (50.55)

and difficulty at birth are the set of statistically significant factors for the nonsusceptible proportion (*long-term effects*) of calves for herd 1, calving difficulty is the only significant factor for herd 7, and there are no significant predictors among this set of covariates for herd 3. We point out that the interpretation of the regression parameters for the cure fraction $\pi(z)$ is such that a higher value for e^{β} would represent a lower probability of cure for the corresponding factor. Note that model (3), together with $\eta(z) = \exp(\beta_\eta z)$ and $\theta(z) = \exp(\beta_\theta z + \beta_c)$, implies that $\pi(z) = (\pi_0)^{e^{\beta}}$, where π_0 represents the probability of cure of the reference group. In particular, calves born in the period march-august have lower probability of cure than those born in september-february; and the probability of cure is a well lower for those that have had difficulties at calving for herd 1. For herd 7 the effect of `difficulty` is the same as for herd 1.

Regarding short-term (mortality) effects, we only find statistically significant predictors in herd 7 where the risk of death of calves born to older mothers, hence with a longer reproductive life, is twice the risk of death of calves born to younger mothers ($\beta_\eta = 0.89$, $e^{\beta_\eta} = 2.44$, p-value $= 0.056$).

Due to a complete parametrization of the probability of cure (survival up to weaning) $\pi(z)$, we can estimate it for each of the categories of the significant covariates for the long-term effects given in Table 2. Table 3 displays the estimate, and confidence interval, for the probability of cure for the different groups for each herd, obtained through the relationship $\pi(z) = (\pi_0)^{e^{\beta}}$. We observe lower probabilities of cure for calves born between March and

TABLE 2. Statistical significant factors for mortality and cure for each herd using a PH-PHC model. Reference group for herd 1: calves born between September and February and without assistance, for herd 7: calves born without assistance.

Predictors	β	e^{β}	se(β)	p	$L_{.95}$	$U_{.95}$
			herd1			
Long term predictor						
`Month`						
mar-aug	1.96	7.097	0.989	0.047	1.022	49.263
`Difficulty`						
slightly assisted	1.87	6.476	0.474	0.000	2.557	16.401
strongly assisted	2.17	8.716	1.051	0.039	1.110	68.386
			herd7			
Long term predictor						
`Difficulty`						
slightly assisted	0.01	1.007	1.027	0.990	0.134	07.549
strongly assisted	1.33	3.798	0.471	0.004	1.507	09.569
Short term predictor						
`Length productive`						
>1300 days	0.89	2.440	0.466	0.056	0.978	06.080

August for herd 1 and for calves born with assistance for herds 1 and 7. Furthermore, note that herd 7 is the only herd for which the length of productive live of the cow has an influence on the risk of death of the calves, and this short-term effect is influencing the probability of cure (survival up to weaning) in such a way that the confidence interval for those calves born with strong assistance $(.887, .953)$ is strictly below the confidence interval for calves born without assistance $(.968, .987)$. Thus,the probability of survival up to weaning of calves born without assistance is significatively higher than the probability of survival up to weaning of calves born with strong assistance.

Concluding, we point out that the PH-PHC model is an alternative to the standard Proportional Hazards model when there is a proportion of non-susceptible individuals in the population. This model allows us to jointly estimate the proportion of cure (survival up to weaning) and the effect of different set of covariates for short and long-term on individuals in a heterogeneous population. Moreover, we have been able to use the same approach for the three different herds, providing a unified method for situations, such as the one described in this paper, where the initial set of covariates has different short-long effects on each herd.

TABLE 3. Estimates of the Probability of Cure $\pi(z)$ and 95% Semiparametric Likelihood Ratio Confidence Intervals (in parentheses). Reference group for herd 1: calves born between September and February and without assistance, for herd 7: calves born without assistance.

Predictors	herd1 ($L_{.95}$, $U_{.95}$)	herd3 ($L_{.95}$, $U_{.95}$)	herd7 ($L_{.95}$, $U_{.95}$)
Reference Group	.993 (.955, .999)	.975 (.955, .986)	.980 (.968, .987)
`Month`			
mar-aug	.953 (.723, .992)		
`Difficulty`			
slightly assisted	.957 (.743, .993)		.981 (.968, .987)
strongly assisted	.942 (.671, .991)		.930 (.887, .953)

Acknowledgments: This research was partially supported by Grants MTM2008-06747-C02-00 and MTM2009-10893 from the Ministerio de Ciencia e Innovación.

References

Farewell, V.T. (1982). The Use of Mixture Models for the Analysis of Survival Data with Long-Term Survivors. *Biometrics*, **38**, 1041-1046.

Goyache, F., Gutiérrez, J.P., Alvarez, I., Fernández, I., Royo, L.J., and Gómez, E. (2003). Genetic analysis of calf survival at different preweaning ages in beef cattle. *Livest. Prod. Sci*, **83**, 13-20.

Sy, J.P., and Taylor, J.M.G. (2000). Estimation in a Cox Proportional Hazards Cure Models. *Biometrics*, **56**, 227-236.

Tarrés, J., Casellas, J., and Piedrafita, J. (2005). Genetic and envieromental factors influencing mortality up to weaning of bruna dels pirineus beef calves in mountain areas. A survival analysis. *Animal Science*, **83**, 543-551.

Tsodikov, A.D. (2002). Semi-parametric models of long- and short-term survival: an application to the analysis of breast cancer survival in Utah by age and stage *Statist. Med.*, **21**, 895-920.

Tsodikov, A.D., Ibrahim, J.G., and Yakovlev, A.Y. (2003a). Estimating Cure Rates From Survival Data: An Alternative to Two-Component Mixture Models. *Journal of the American Statistical Association*, **98**, 464, 1063-1078.

Tsodikov, A.D. (2003b). Semiparametric models: a generalized self-consistency approach. *Journal Royal Statistical Society B*, **65**, Part 3, 759-774.

Semi-parametric Modelling for Extremes with Threshold Estimation

Anna MacDonald[1], Carl Scarrott[1], Dominic Lee[1]

[1] Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand (carl.scarrott@canterbury.ac.nz)

Abstract: Extreme value modelling is used to characterise the properties of unusual or rare events, e.g. the upper or lower tails of a distribution. A semi-parametric mixture model is proposed which combines a smooth non-parametric density estimator with an asymptotically motivated parametric tail model. The threshold which switches between the bulk of the distribution, and the tail is automatically estimated within the inference procedure, overcoming the obstacle of threshold choice and uncertainty estimation in traditional extremal tail modelling approaches. Bayesian inference is used to account for all uncertainties. The model is demonstrated for quantile estimation of daily FTSE100 log returns.

Keywords: Extremal Mixture Modelling; Bayesian Inference; Kernel Density.

1 Introduction

The classical models derived using extreme value theory are used to approximate properties of the unusual behaviour of a process, as opposed to the 'average' behaviour that most statistical models target. The generalised Pareto distribution (GPD) is one such model, typically used to approximate the distribution of the exceedances over some suitably high threshold u. The exceedances denoted by $X_1, \ldots, X_n$ can often be assumed to follow a GPD(ϕ_u, σ_u, ξ) such that for $x > u$:

$$\Pr(X > x) = \begin{cases} \phi_u \left[1 + \xi \left(\dfrac{x-u}{\sigma_u}\right)\right]_+^{-1/\xi} & \xi \neq 0 \\ \phi_u \exp\left[-\left(\dfrac{x-u}{\sigma_u}\right)\right]_+ & \xi = 0, \end{cases}$$

where $[\cdot]_+ = \max(\cdot, 0)$, ξ and $\sigma_u > 0$ are the shape and scale parameters respectively and the final parameter is $\phi_u = \Pr(X > u)$. The value of the shape parameter ξ determines the limiting tail behaviour; $\xi = 0$ results in an exponential tail, $\xi < 0$ gives a tail with finite support ($0 < x - u < -\sigma_u/\xi$) and $\xi > 0$ has a heavier than exponential tail with no upper-end point, see Coles (2001) and references therein for further details. The dependence of the scale σ_u on the threshold u is discussed further below.

A challenge in the application of the GPD is the choice of a 'suitably high threshold'. In some applications the threshold choice has a strong impact on tail extrapolations. We wish to choose a sufficiently low threshold to reduce the variance of our sample estimates, without going so low to invalidate the asymptotic arguments used to motivate the GPD. Typically, graphical diagnostics are used to assess model fit for a range of thresholds, but unfortunately these often require subjective expert judgement. Further, once the threshold is chosen it is typically treated as a fixed quantity, so the associated uncertainty is not accounted for.

Much recent research effort has considered mixture models to overcome the issues surrounding threshold selection. These mixtures attempt to describe simultaneously the bulk of the distribution with an extreme value model for the upper or lower tail. Behrens *et al.* (2004) combine a parametric density for the bulk below an estimated threshold and the GPD above it. Tancredi *et al.* (2006) consider piecewise uniform distributions from a very low threshold up to the estimated threshold and an extreme value model for excesses above this threshold. Frigessi *et al.* (2002) present a two component dynamically-weighted mixture model with a transition function between the bulk and tail model.

We have developed a new semi-parametric mixture model which combines a non-parametric density estimator for the distribution below the threshold and an extreme value model above the threshold. Our proposed model overcomes the need to specify a parametric form for the bulk distribution as in most of the aforementioned approaches, and provides a very flexible smooth model for the bulk of the distribution.

2 Proposed Mixture Model

Consider $\mathbf{X} = \{X_1, ..., X_n\}$ a sequence of iid observations. The proposed mixture model has density function for a single observation given by:

$$f(x|\theta, \mathbf{X}) = \begin{cases} (1 - \phi_u) \cdot h(x|\lambda, \mathbf{X}) & x < u \\ \\ \phi_u \cdot g(x|\xi, \sigma_u, u) & x \geq u \end{cases}$$

where $g(\cdot)$ is the density function for the GPD discussed above and $h(\cdot)$ is the usual univariate kernel density estimator (for our application using a Gaussian kernel) and $\lambda > 0$ controls the smoothness of the density estimate. The parameter vector is $\theta = (\lambda, \xi, \sigma_u, u)$ and $\phi_u = \Pr(X > u)$ which is estimated using the sample proportion. We will use the cross-validation likelihood given in Brewer (2000) to estimate the bandwidth λ within a Bayesian inference approach below.

An alternative representation of the GPD is available which removes the dependence of the parameters on the threshold making specification of priors straightforward and leads to better mixing of the chains in the Bayesian

inference approach outlined in the following Section. It turns out that for a sufficiently high threshold u the point process defined by $(i/n, X_i)$ for $i = 1, \ldots, n$ can be well approximated by a non-homogeneous Poisson process on the region $(0,1)\times[u,\infty)$, where the intensity function on a subregion $B = (t_1, t_2) \times (x, \infty)$ is given by:

$$\Lambda(B) = (t_2 - t_1)n_b \left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi} \text{ for } x > u,$$

the scaling constant n_b is seen as completely arbitrary and is set to be the number of threshold exceedances, which is known to improve the likelihood condition (and hence chain mixing properties below), see Coles (2001) for further details. The GPD can be derived as a special case of this point process representation, where the parameters (μ, σ, ξ) are not dependent on the threshold with $\sigma_u = \sigma - \xi(u - \mu)$.

3 Bayesian Inference

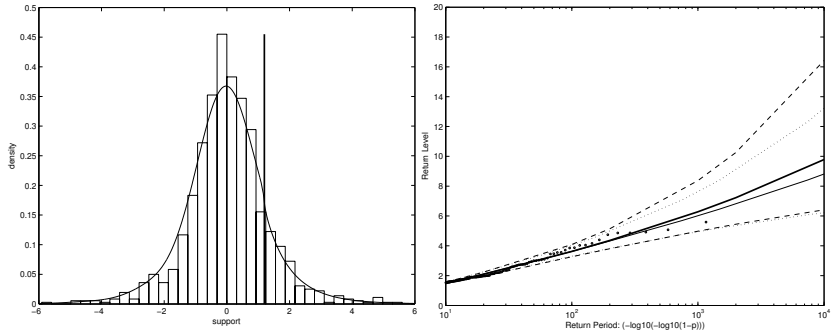
Computation for the posterior distribution is achieved via Markov chain Monte Carlo (MCMC), with a Metropolis-Hastings random walk sampler being used. Using the point process representation the mixture model parameter vector is $\theta = (\lambda, u, \xi, \sigma, \mu)$. The joint prior distribution is factorised to be $\pi(\lambda, u, \xi, \sigma, \mu) = \pi(\lambda) \cdot \pi(u) \cdot \pi(\xi, \sigma, \mu)$, with the marginals given by:

- $\pi(\lambda)$ - Inverse Gamma$(\lambda^2|d_1, d_2)$ as proposed by Brewer (2000).
- $\pi(u)$ - $N(u|\mu_u, \nu_u^2)$, where μ_u is set at a high quantile with large ν_u^2 to give diffuse prior.
- $\pi(\xi, \sigma, \mu)$ - MVN(μ_{pp},Σ), with Σ the covariance structure representing independence for a naive analysis.

Convergence of the chains is assessed by standard diagnostic procedures.

4 Application

The FTSE100 is a share index of the 100 most highly capitalised blue chip companies listed on the UK market. We demonstrate the proposed mixture model by application to the distribution of log returns on the closing price of the FTSE100 from 01/06/00 to 31/12/04 ($n = 1159$) shown in Figure 1(a). This application is for demonstration purposes only, as it is well known that financial returns exhibit various forms of dependence and possible non-stationarity. However, we are only interested in the performance of our model in approximating a marginal density for the financial returns. Figure 1(a) shows the data histogram with posterior predictive density estimate and estimated threshold. The nonparametric density estimate clearly



(a) *Histogram with posterior predictive density.*

(b) *Return level plot with 90% credible intervals*

FIGURE 1. Results for FTSE100 log returns.

provides an adequate fit. The tail fit is typically assessed using a return level plot (quantiles against cumulative probability on a negative log-log scale), see Coles (2001), as shown in Figure 1(b). For comparison the posterior predictive return levels for both the mixture model (bold and dashed lines) and for the point process model only (thin and dotted lines) are given. The threshold for the point process was set to the same value as for the mixture model, which is close to what would have been chosen using standard graphical diagnostics. It is clear from the return level plot that the mixture model provides a good fit (and realistic extrapolation) in the upper tail. Further, the wider credible intervals for the mixture model (compared to the traditional fixed threshold approach) demonstrate the ability of the method to account for the extra uncertainty due to threshold choice.

References

Behrens, C.N., Lopes, H.F., and Gamerman, D. (2004). Bayesian analysis of extreme events with threshold selection. *Statistical Modelling*, **4**, 227-244.

Brewer, M.J. (2000). A Bayesian model for local smoothing in kernel density estimation. *Statistics and Computing*, **10**, 299-309.

Coles, S. (2001). *An introduction to statistical modelling of extreme values*. London: Springer-Verlag.

Frigessi, A., Haug, O., and Rue, H. (2002). A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes*, **5**, 219-235.

Tancredi, A., Anderson, C., and O'Hagan, A. (2006). Accounting for threshold uncertainty in extreme value estimation. *Extremes*, **9**, 87-106.

Interval censored PH survival models for longitudinal data: precision of estimators

Gilbert MacKenzie[1], Defen Peng[2]

[1] ENSAI, France & Centre of Biostatistics, University of Limerick, Ireland
[2] Centre of Biostatistics, University of Limerick, Ireland & Department of Statistics, Zhongnan University of Economics and Law, China

Abstract: We present a general likelihood for interval censored survival data arising in longitudinal studies such as longitudinal randomized controlled clinical trials (RCTs) and give some general formulae for inference in parametric, interval censored, proportional hazards, regression survival models. For the exponential regression model we compare the performance of the general likelihood with a commonly used proxy likelihood, which ignores the interval censoring by treating the interval censored times to events as if they were exact. We show analytically that use of the proxy likelihood leads to estimators (for example, of the treatment effect) which are artificially precise and we quantify the extent of the resulting biases in a simulation study.

Keywords: Artificial precision, Interval Censoring, Longitudinal RCTs; PH Survival Models, Proxy likelihood.

1 Introduction

We consider the longitudinal randomized controlled clinical trials (LDA-RCT setting, where the response variable, $Y(t)$, is binary). Typically at baseline (t_0) the ith patient is in healthy state, i.e., $Y_i(t_0) = 0$, and as the process evolves an adverse event may occur, i.e., $Y_i(t_s) = 1$ where $t_s > t_0$. Finkelstien (1986) and Collett (1994) elected to adopt a "time to event" analysis in order to recover information on the treatment effect in the LDA-RCT setting. Moreover, clinicians (Bergink *et al.*, 1998) have adopted a similar approach in which interval censored follow-up times, to the loss of 3 lines of visual acuity (Bailey-Lovie, 1976), were treated as if they were exact times to events. Intuitively, this simple expedient which is commonly adopted in applied settings seems sub-optimal and the objective of this note is to investigate the extent of any penalty incurred.

In this context, Finkelstien (1986) argued that a common fixed follow-up examination schedule was required for each individual and this approach was echoed by Collett (1994) following a multinominal scheme suggested by Lawless (1974). However, typically, it is not respected by individuals in

clinical trials. Accordingly, in the course of our development, we, *inter alia*, relax this rather un-necessarily restrictive assumption.
The paper is organized as follows. In §2 we develop the general interval-censored likelihood in the LDA setting and discuss the usual form of the proxy likelihood. In §3 we give a motivating example and in §4 develop the Exponential Regression model giving key results. Lastly, in §5 we remark on some extensions.

2 Likelihoods

Suppose there are $m+1$ *fixed*, scheduled, inspection times, t_0^*, t_1^*,..., t_m^* at which continuous or ordinal responses Y_0, Y_1,..., Y_m, are measured. This arrangement implies m+1 time intervals: $I_1 = (t_0, t_1^*], I_2 = (t_1^*, t_2^*], ..., I_k = (t_{k-1}^*, t_k^*], ..., I_m = (t_{m-1}^*, t_m^*]$ and $I_{m+1} = (t_m^*, \infty]$. Typically, $t_0 = 0$, especially in RCTs where, $t_0 = 0$ represents time of randomization. Hence, let T be a non-negative random variable denoting the time to some outcome of interest defined on the Ys. Let $S(t;\theta)$ and $\lambda(t;\theta)$ be the corresponding survival and hazard functions, respectively, depending on the unknown possibly vector-valued parameter $\theta \in \Theta$. Then, for a sample of n independent subjects subject to non-informative censoring the usual likelihood for the unknown parameters is

$$L(\theta) = \Pi_{i=1}^{n}[\lambda(t_i;\theta)S(t_i;\theta)]^{\delta_i}[S(t_{ic};\theta)]^{1-\delta_i}, \tag{1}$$

where $\lambda(t_i;\theta)S(t_i;\theta) = f(t_i;\theta)$, δ_i is the censoring indicator ($\delta_i = 1$ for an event and 0 otherwise) and t_{ic} is a right censored survival time.
Typically each individual ($i = 1, \ldots, n$) defines their own trajectory over the course of the longitudinal study, thereby generating a person-specific set of intervals. Accordingly, we obtain the following interval censored likelihood,

$$L_1(\theta) = \Pi_{i=1}^{n}[S(t_{i,k-1};\theta) - S(t_{ik};\theta)]^{\delta_i}[S(t_{ic};\theta)]^{1-\delta_i}, \tag{2}$$

where the actual times at which the ith patient presents for examination are utilized in the likelihood. Typically, t_{ik} is close to t_k^*, but this is not always the case.
Overall, conditioning on the times observed is to be preferred as it obviates the incorporation of false assumptions about arrival and departure times in the likelihood for the interval-censored observations. It is convenient to rewrite (2) as

$$L_1(\theta) = \Pi_{i=1}^{n}\{S(t_{i,k-1};\theta)[1 - S(t_{i,k-1}, t_{ik};\theta)]\}^{\delta_i}[S(t_{ic};\theta)]^{1-\delta_i}. \tag{3}$$

Having obtained a realistic interval censored likelihood for longitudinal RCTs, we turn to consider proxy likelihoods arising from the *ad-hoc* approaches which are currently in use. The main approach is simply to substitute directly one of: (a) the beginning point of the interval, t_{ib}, or (b) the

interval mid-point, t_{im} or, (c) the interval end-point, $t_{ie}, \forall\, i$, as if it were the exact time at which failure occurred in (1) above. Thus, when substituting one of these inexact times we may use (1) directly as a proxy-likelihood denoting it then by $L_2(\theta)$.
Now, $L_1(\theta)$ and $L_2(\theta)$ may be used to compare inference in a survival model and estimate the penalty, if any, associated with treating interval censored survival observations as if there were exact times to events.

3 A Motivating Example

If T follows the Exponential distribution with parameter ϕ, then $\lambda(t;\theta)=\phi$, $S(t;\theta)=\exp(-\phi t)$ and $S(t_{i,k-1}, t_{ik};\theta)= \exp[-\phi(t_{ik} - t_{i,k-1})]$. From (1) and writing $\ell_2(\phi)$ for $\log_e L_2(\phi)$, the first derivative is given by

$$U_2(\phi) = \frac{\partial l_2(\phi)}{\partial \phi} = \sum_{i=1}^{n} [\delta_i \phi^{-1} - \delta_i t_i - (1-\delta_i) t_{ic}] \tag{4}$$

and solving $U_2(\phi) = 0$ yields the closed form MLE

$$\hat{\phi} = \frac{n_u}{(T_u + T_c)}, \tag{5}$$

where T_u and T_c are the sums of the uncensored and censored times respectively, and $n_u = \sum_{i=1}^{n} \delta_i$ is the total number of uncensored events. Differentiating again we find

$$I_2(\phi) = -\frac{\partial^2 l_2(\phi)}{\partial \phi^2} = \sum_{i=1}^{n} \delta_i \phi^{-2}. \tag{6}$$

Thus, the variance of $\hat{\phi}$ is given by $V_2(\phi) = \phi^2/n_u$ which may be consistently estimated by substituting $\hat{\phi}$.
The corresponding equations for (4) are

$$U_1(\phi) = \sum_{i=1}^{n} [\delta_i d_i(t_k) \omega_i(\phi) - \delta_i t_{i,k-1} - (1-\delta_i) t_{ic}], \tag{7}$$

where $d_i(t_k) = (t_{ik} - t_{i,k-1})$ and $\omega_i(\phi) = S(t_{i,k-1}, t_{ik};\phi)/[1 - S(t_{i,k-1}, t_{ik};\phi)]$ the conditional odds on survival in the interval $(t_{i,k-1}, t_{ik}]$. Since $\omega_i(.)$ is non-linear in ϕ, the ML estimating equation, $U_1(\phi) = 0$, must be solved iteratively for $\hat{\phi}$. However, it may be shown that the MLE given by (5) is the approximate (i.e., first order) solution of $U_1(\phi) = 0$. The observed information is then

$$I_1(\phi) \approx \sum_{i=1}^{n} \delta_i d_i^2(t_k) \omega_i(\phi)[1 + \omega_i(\phi)] \tag{8}$$

and the approximate asymptotic variance of $\hat{\phi}$ is given by $V_1(\phi) \approx \phi^2/(n_u - \phi d)$ where $d = \sum_{i=1}^{n} \delta_i d_i(t_k)$. We may compare the relative efficiency of the two estimators by examining $V_2(\phi)/V_1(\phi) = 1 - (\phi d/n_u) < 1$, which shows that the estimator based on $L_2(\theta)$ under-estimates the true variance $V_1(\phi)$ when the observed inspection times are analyzed as if they were exact. We may gain further insight by substituting $\hat{\phi}$ from (5) to show that the relative efficiency is approximately $(T_u + T_c - d)/(T_u + T_c)$, a factor which artificially increases the precision of the estimator based on (1), as the time intervals between visits coarsen (MacKenzie, 1999). This is an approximate result, but it has been confirmed in simulation studies (Peng, 2009).
This finding leads to the conjecture, arguably rather bold, that a similar result holds, i.e., $V_2(\phi)/V_1(\phi) < 1$, for all PH models.

4 The Exponential Regression Model

4.1 Proxy Likelihood

Armed with these general formulae we investigate the Exponential Regression model. Let T follow the exponential regression model defined by

$$\lambda_{i2} = \lambda(t_i; \alpha_2, \beta_2) = \exp(\alpha_2 + x_i'\beta_2),$$

where $S(t_i; \alpha_2, \beta_2) = \exp[-\lambda_{i2} t_i]$ and α_2 is an unconstrained parameter, β_2 is $p \times 1$ vector of regression coefficients and x_i is a $p \times 1$ vector of fixed covariates. The corresponding proxy likelihood is

$$L_2(\alpha_2, \beta_2) \quad = \quad \prod_{i=1}^{n} \left\{\lambda_{i2} \mathrm{e}^{-\lambda_{i2} t_i}\right\}^{\delta_i} \left\{\mathrm{e}^{-\lambda_{i2} t_{ic}}\right\}^{1-\delta_i}, \tag{9}$$

4.2 IC likelihood

For the IC likelihood we have

$$\lambda_{i1} = \lambda(t_i; \alpha_1, \beta_1) = \exp(\alpha_1 + x_i'\beta_1),$$

where $S(t_{i,k-1}, t_{ik}; \alpha_1, \beta_1) = \exp[-\lambda_{i1} d_i(t_k)]$, and $d_i(t_k) = t_{ik} - t_{i,k-1}$ is the width of the kth interval. Then,

$$L_1(\alpha_1, \beta_1) = \prod_{i=1}^{n} \left\{\mathrm{e}^{-\lambda_{i1} t_{i,k-1}} \left[1 - \mathrm{e}^{-\lambda_{i1} d_i(t_k)}\right]\right\}^{\delta_i} \left\{\mathrm{e}^{-\lambda_{i1} t_{ic}}\right\}^{1-\delta_i}, \tag{10}$$

4.3 Comparison of IC and Proxy Approaches

Comparing the Proxy and IC approaches we find that approximate IC mles are identical to those estimated at $t_{ie} = t_{ik}$, the end points of the interval using the proxy likelihood (ie, $\hat{\alpha}_1 = \hat{\alpha}_2$ and $\hat{\beta}_{1r} = \hat{\beta}_{2r}$) with proxy t_{ie}.

We compared the relative efficiency of the two estimators by examining $V_2(\hat{\alpha}_2)/V_1(\hat{\alpha}_1)$ and $V_2(\hat{\beta}_{2r})/V_1(\hat{\beta}_{1r})$, $r = 1, 2, \cdots, p$. The details are too lengthy to reproduce here. Analytical results are available only for categorical covariates. We have proved the following result for a categorical covariate with $p+1$ categories, modelled by p binary dummy variables, i.e.

$$\begin{aligned} V_2(\hat{\alpha}_{2e})/V_1(\hat{\alpha}_1) &< 1 \\ V_2(\hat{\beta}_{2er})/V_1(\hat{\beta}_{1r}) &< 1 \end{aligned} \tag{11}$$

so that the conjecture that the proxy mles are artificially precise holds, under the first order conditions invoked above, for categorical covariates. In the journal paper we confirm these findings in a simulation study using the full IC likelihood and for continuous covariates. We note in passing that any continuous covariate may be represented in $p \leq n$ distinct categories and hence for such a representation of a continuous covariate the above conjecture holds.

5 Summary

In the journal paper we extend these ideas by developing inference for the general interval-censored likelihood in the parametric PH case and extend the methods to other PH regression models. In subsequent sections we conduct a simulation study which confirms our analytical findings that the IC likelihood is the most appropriate vehicle for inference. We also analyze data from the MRC's multi-centre RCT of Teletherapy in ARMD (Hart *et al*, 2002) and consider extensions to non-PH models. Elements of these findings will be presented at the workshop.

Acknowledgments: The work was supported by the Science Foundation Ireland (SFI, www.sfi.ie) Mathematics Initiative, II, via the BIO-SI (www.ul.ie/bio-si) research programme in the Centre of Biostatistics, University of Limerick, Ireland: grant number 07/MI/012. In addition, Professor Peng is also supported by SFI via a Research Frontiers Programme award, grant number 05/RF/MAT 026.

References

Bailey IL, Lovie JE (1976). New design principles for visual acuity letter charts. Am J Optom Physiol Opt.53:740745.

Bergink GJ, *et al* (1998). A Randomised Controlled Clinical Trial on the efficacy of radiation therapy in the control of subfoveal choroidal noevascularisation in age-related macular degeneration: radiation versus observation. *Graefe's Arch. Clin. Exp.Ophthalmology.* 236, No 752, 1-5.

Collett D (2008). *Modelling survival data in medical research.* CRC, London.

Hart P *et al* (2002). Visual outcomes in the subfoveal radiotherapy study. Arch Ophthalmol. 120:1029-1038.

Finkelstein D (1986). A Proportional Hazards Model for Interval-Censored Failure Time Data *Biometrics.* 4, 2, 845-854.

Lawless J (1974). *Statistical Models and Methods for Lifetime Data* Wiley, New York.

MacKenzie G (1999). Survival analysis for longitudinal data. Proceedings of the 14th International Workshop on Statistical Modelling, July, Graz, Austria. July 1999, pages 259-264.

Macular Photocoagulation Study Group (1994). Visual outcome after laser photocoagulation for sub-foveal choroidal neovascular secondary to age-related macular degeneration. *Arch. Ophthalmol.* 112, 480-488.

Peng D (2009). *Inferences in the Interval Censored Exponential Regression Model.* Masters Thesis MacMaster University, Ontario, Canada.

Spatiotemporal Modelling of Nitrate and Phosphorus in River Catchments for England and Wales

Ana-Maria Magdalina[1], Claire Ferguson[1], Adrian Bowman[1], Robert Willows[2], David Johnson[2], Chris Burgess[2], Linda Pope[2], Marian Scott[1], Duncan Lee[1]

[1] Department of Statistics, 15 University Gardens, University of Glasgow
[2] Environment Agency, Evidence Directorate

Abstract: The Environment Agency has monitored river water quality in England and Wales over the last 40 years. In order to investigate the impact and implications of European directives for surface water, it is of interest to analyze the overall trends in the level of nutrient concentrations present in such rivers and understand their spatiotemporal distribution. We are specifically interested in changes in nutrients across time and space as a consequence of environmental factors, agricultural practice, population and river catchment characteristics.

Keywords: Spatiotemporal modelling; Environmental statistics; River catchments.

1 Introduction

The Environment Agency has more than 6000 river monitoring locations in England and Wales that are sampled each year on a monthly basis. These sites are contained in small waterbodies, which are the standard surface water reporting units for the Water Framework Directive (WFD, *European Parliament*, 2000). A collection of waterbodies, covering a river network, define a large hydrological area, of which there are 59 in England and Wales. In order to investigate the impact and implications of directives such as the WFD and the Nitrates Directive (*Nitrates Directive*, 1991), it is of interest to examine the overall trends in the level of nutrients present in rivers and understand their spatiotemporal distribution.

This paper presents the analysis for monitoring locations on the River Soar. The River Soar is located in the English East Midlands and it is a tributary of the River Trent, which is one of the main rivers in England.

The River Soar area contains measurements that span the period 1985 - 2009. However, only a small percentage of the 177 monitoring locations have measurements for the full period. Moreover, as is often the case with environmental data, some of these measurements are less-than values (left

censored), and the limits of detection vary across time. For less than 3% censored the detection limit is taken as the observed value, between 3% and 50% censored, imputation methods have been applied ($Helsel, D.R., 2005$) and sites with more than 50% censored data are not included.
Across the 177 monitoring locations in the Soar region, there are measurements for various forms of phosphorus and nitrate. For illustrative purposes only the orthophosphate determinand will be discussed in this paper.

2 Methods

An additive model is fitted for each determinand, orthophosphate in this case, in which the response variable is expressed as a sum of smooth functions of the space and time components:

$$y = \mu + s(\text{Easting, Northing}) + s(\text{year.day}) + s(\text{doy}) + \varepsilon \tag{1}$$

where *doy* represents the day of year and $year.day = year + \frac{doy}{366}$.
The smooth functions are constructed using local linear regression, with one and two covariates for time and space respectively. The weights, for smoothing, are based on a normal density with standard deviation, h, where h is the smoothing parameter automatically chosen using 6 degrees of freedom for univariate terms and 12 for bivariate. For *doy* a cyclic smoother was applied with local mean regression, for full details see Bowman et al. (2009). Initially, this model has been fitted assuming additive effects i.e. without investigating if there is a space-time interaction and the errors are assumed to be independent, $\varepsilon \sim N(0, \sigma^2)$. However, interactions and correlations can be incorporated in the modelling as described in Bowman et al. (2009). The model fits a smooth trend for space, a smooth trend for time and a smooth trend for seasonality.

3 Results

The model in equation (1) requires sufficient data both in time and across space and therefore we have selected for analysis only those sites on the River Soar that had a minimum of 100 observations in the last 15 years. Therefore, for orthophosphate 99 sites were selected that for the 1995-2009 period each contained more than 100 observations.
Exploratory work highlights that a log transform is necessary to stabilize the variance over time and lessen the impact of outliers. Overall, the level of orthophosphate appears to decrease with time and the spatial pattern highlights large values in the South where the source of the river lies and also in the North, on the last section of the River Soar, before it becomes a tributary of the River Trent. The trends were not typically linear and there

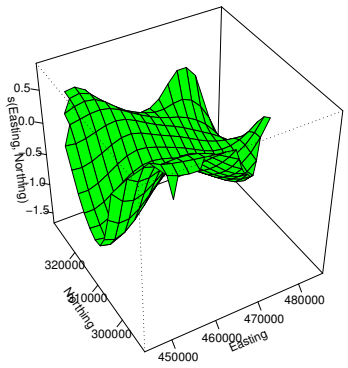


FIGURE 1. The 3D surface plot shows the spatial component of the additive model. The x-axis represents Easting, z the Northing and y the smoothed spatial trend from the additive model.

is evidence of a seasonal pattern with higher values in the summer. Model 1 was fitted using the data described above and the results are shown in Figures 1 and 2.

Figure 1 highlights the areas with higher levels of orthophosphate. Values appear large at high and low values of Northing and low on the East and West most boundaries. It should be noted that there are less measurements available around the high and low Easting coordinates.

Figure 2 (left) highlights the year component after fitting the additive model and indicates a downward long term trend suggesting a reduction in phosphorus and an improvement in water quality. Figure 2 (right) highlights the plot of the day of the year component from the additive model and indicates a cyclic pattern with higher values in the summer.

Present work is focused on extending the additive model to explain the variability in the nutrient data by including further covariates such as population, rainfall, land use, land cover, physical characteristics of the land and river characteristics.

4 Future work

For each large hydrological area a separate spatio-temporal model will be fitted to investigate how much of the variation can be explained by the

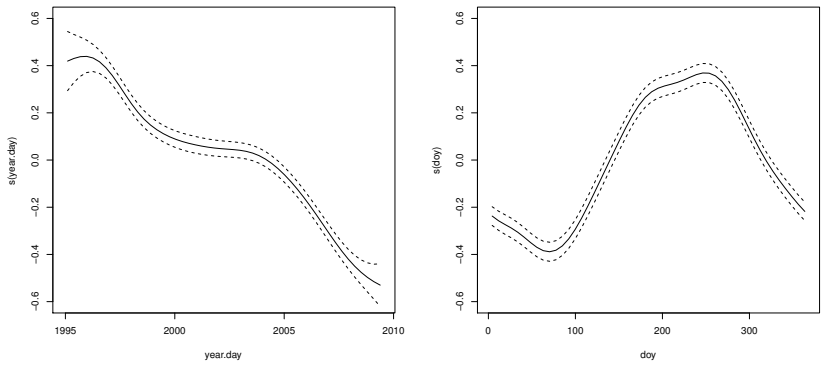


FIGURE 2. The left panel shows the year component and the right panel displays the day of the year components from the additive model. The solid lines indicate the estimates for the trend and the seasonal pattern respectively, with +/− 2 standard errors represented by dashed lines.

covariates. To incorporate the covariates the present modelling will be developed at waterbody level for areal data.

Acknowledgments: AMM gratefully acknowledges funding from the Environment Agency and the University of Glasgow for this work and is thankful to the Environment Agency for data access and support.

References

Bowman, A.W., Giannitrapani, M. and Scott, E.M. (2009). Spatiotemporal smoothing and sulphur dioxide trends over Europe. *Journal of the Royal Statistical Society, Series C*, **58** (5), 737-752.

Helsel, D.R. (2005). *Nondetects and data analysis: statistics for censored environmental data*. New York: Wiley-Interscience.

European Parliament (2000). Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000, establishing a framework for community action in the field of water policy. *Official Journal of the European Communities*, **L327/1**, 1-72.

Nitrates Directive (1991). COUNCIL DIRECTIVE of 12 December 1991 concerning the protection of waters against pollution caused by nitrates from agricultural sources (91/676/EEC).

Multiple change-point identification in models valid for describing trends of disease incidence or mortality rates

Nirian Martin[1], Leandro Pardo[2]

[1] Department of Statistics, Carlos III University of Madrid, Spain
[2] Department of Statistics and O.R., Complutense University of Madrid, Spain

Abstract: In this paper the problem of change-point identification in models that allow describing trends for incidence or mortality rates is studied on the basis of an underlying Generalized Poisson regression model with linear constraints. The Annual Percent Change (APC) is an important statistical measure for describing trends in rates that is related to the aforementioned model. In order to detect the position where the trends have changed along the time-line, we propose a hierarchical sequence of nested hypotheses and its corresponding maximum likelihood estimators and likelihood ratio test-statistics.

Keywords: Annual Percent Change (APC); Poisson sampling; Change-point.

1 Introduction

In order to implement prevention and control plans against diseases, it is very important for health authorities to analyze whether the incidence or mortality rates related to such diseases have increased or decreased in a set of successive years and in what degree, that is the trends in incidence or mortality are taken into account. For studying trends in rates of several disease such as HIV/AIDS, cancer or myocardial infarction, the Annual Percent Change (APC) has been considered, see for instance Hall et al. (2005), Weir et al. (2003), Rosamond et al. (1998). Recently, new techniques that are needed to cover different problems related to the age-adjusted disease rates are being continuously implemented. For instance in Li et al. (2008) an age-stratified Poisson regression model was proposed for constant trends, focussed on cancer rates from the National Surveillance, Epidemiology and End Results (SEER) Program of the National Cancer Institute (NCI) database. We shall consider J age-groups in the study and because it is supposed constant trend only inside K subperiods. The end-point and starting-point of $K-1$ consecutive subperiods are the change-points that we want to detect through likelihood based statistical techniques. In each subperiod k, $k = 1, ..., K-1$, we have a sequence of I_k time points$\{t_{ki}\}_{i=1}^{I_k}$, and independent Poisson random variables that count

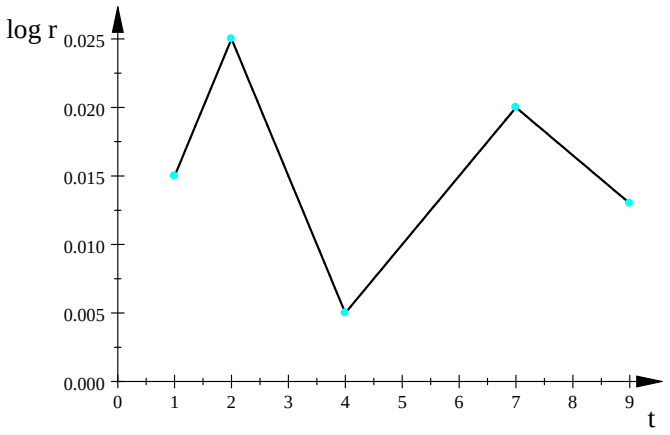


FIGURE 1. $\log r_{ki} = \beta_{0k} + \beta_{1k} t_{ki}$ for $\{t_{ki}\}_{i=1}^{I_k}$, $k = 1, ..., K$, with $K = 4$, $I_1 = 2$, $I_2 = 2$, $I_3 = 3$, $I_4 = 2$.

incidence or mortality events are taken into account, $D_{kji} \overset{\text{ind}}{\sim} \mathcal{P}(m_{kji}(\boldsymbol{\beta}))$, $j = 1, ..., J$, $i = 1, ..., I_k$, $\boldsymbol{\beta} = (\beta_{01}, ..., \beta_{0J}, \beta_{11}, ..., \beta_{1K})^T \in \mathbb{R}^{J+K}$. The unknown parameter associated with subperiod k, follows a Generalized Poisson Regression model, defined by

$$\log \frac{m_{kji}(\boldsymbol{\beta})}{n_{kji}} = \beta_{0j} + \sum_{h=1}^{k-1} \beta_{1h}(t_{hI_h} - t_{h-1,I_{h-1}}) + \beta_{1k}(t_{ki} - t_{k-1,I_{k-1}}), \quad (1)$$

$k = 1, ..., K$, where n_{kji} is the population at risk in the k-th subperiod (known values). We can define the expected stardardized incidence or mortality ratio at time-point i as

$$\begin{aligned} r_{ki} &= \textstyle\sum_{j=1}^{J} \omega_j \dfrac{m_{kji}(\boldsymbol{\beta})}{n_{kji}} \\ &= \exp(\beta_0) \exp\left(\sum_{h=1}^{k-1} \beta_{1h}(t_{hI_h} - t_{h-1,I_{h-1}}) + \beta_{1k}(t_{ki} - t_{k-1,I_{k-1}}) \right), \end{aligned}$$

where $\exp(\beta_0) \equiv \sum_{j=1}^{J} \omega_j \exp(\beta_{0j})$ and $\{\omega_j\}_{j=1}^{J}$ is the age-distribution of the Standard Population ($\sum_{j=1}^{J} \omega_j = 1$, $\omega_j > 0$, $j = 1, ..., J$). The APC in the k-th subperiod is defined as $\text{APC}_k = 100(\exp(\beta_{1k}) - 1)$, and represents an average rate of change per year in a given period of time framework when constant change along the time has been assumed.

In order to establish the sequence of nested *Change-point Poisson Regression Models*, the linear constraints constitute the way to joint two consecutive intervals or to remove the change-point between both. If there is no

change-point at t_{sI_s}, then for the $(s+1)$-th subperiod, $s \in \{1, ..., K-1\}$, it should be held $\beta_{1s} - \beta_{1,s+1} = 0$. We are going to build the parameter space by agregation: let $t_{s_2 I_{s_2}}$ the first change point that we shall remove, then $\Theta_{K-1} = \{\boldsymbol{\beta} \in \mathbb{R}^{J+K} : \beta_{1s_1} - \beta_{1,s_1+1} = 0\}$; let $t_{s_2 I_{s_2}}$ the second change point that we shall remove, then $\Theta_{K-2} = \Theta_{K-1} \cap \{\boldsymbol{\beta} \in \mathbb{R}^{J+K} : \beta_{1s_2} - \beta_{1,s_2+1} = 0\}$; ...; let $t_{s_r I_{s_r}}$ the r-th change point that we shall remove, then the parameter space is

$$\Theta_{K-r} = \Theta_{K-(r-1)} \cap \{\boldsymbol{\beta} \in \mathbb{R}^{J+K} : \beta_{1s_r} - \beta_{1,s_r+1} = 0\}. \tag{2}$$

In the literature for change-point inference the most common distribution related to the underlying model is the continuous one and discrete probability models are usually avoided. Even though for incidence and mortality counts Poisson assumption is theoretically assumed, later in practice an underlying Normal distribution is followed (see for instance Kim et al. (2000)). For change-point detection several methods can be encountered, for instance non-parametric methods (Kim et al. (2000)) and Bayesian methods (Tiwari et al. (2005)) and our proposed methodology is based on the likelihood ratio test. More precisely, based on the likelihood ratio test binary segmentation technique can be used (see for instance, Worsley (1983)) but we propose a sequence test-statistics based on sequence of nested hypotheses. In the sequence of K nested hypotheses

$$\mathcal{H}_0(k) : \boldsymbol{\beta} \in \Theta_k \quad \text{vs} \quad \mathcal{H}_1(k) : \boldsymbol{\beta} \in \Theta_{k+1} - \Theta_k, \quad k = 0, ..., K-1, \tag{3}$$

where $\Theta_0 \subset \Theta_1 \subset \cdots \subset \Theta_K$. Subscript k in the null hypothesis means that the Poisson regression model has k change points and the alternative hypothesis means that we have $k+1$ change points and in addition it is not possible to remove one.

2 Maximum likelihood estimators

Based on the likelihood function of a Poisson sample $\boldsymbol{D} = (\boldsymbol{D}_1, ..., \boldsymbol{D}_K)^T$, with $\boldsymbol{D}_k = (D_{k11}, ..., D_{k1I_k}, ..., D_{kJ1}, ..., D_{kJI_k})^T$, $k = 1, ..., K$, the the kernel of the log-likelihood function is given by

$$\ell_{\boldsymbol{\beta}}(\boldsymbol{D}) = \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I_k} D_{kji} \log m_{kji}(\boldsymbol{\beta}) - \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I_k} m_{kji}(\boldsymbol{\beta}).$$

The MLE of $\boldsymbol{\beta}$ under $\mathcal{H}_0(K-r)$ is

$$\widehat{\boldsymbol{\beta}}_{\mathcal{H}_0(K-r)} = \arg \max_{\boldsymbol{\beta} \in \Theta_{K-r}} \ell_{\boldsymbol{\beta}}(\boldsymbol{D}) = \arg \max \left\{ \max_{\boldsymbol{\beta} \in \Theta_{K-r}(h)} \ell_{\boldsymbol{\beta}}(\boldsymbol{D}) \right\}_{h=1}^{K-(r-1)},$$

where $\Theta_{K-r}(h) = \Theta_{K-(r-1)} \cap \{\boldsymbol{\beta} \in \mathbb{R}^{J+K} : \beta_{1q_h} - \beta_{1,q_h+1} = 0\}$,$r = 1, 2, ..., K$, $\Theta_K = \mathbb{R}^{J+K}$, $\{t_{q_h I_{q_h}}\}_{h=1}^{K-(r-1)} = \{t_{kI_k}\}_{k=1}^{K} - \{t_{s_h I_{s_h}}\}_{h=1}^{r-1}$ is the

set of change points not removed according to $\Theta_{K-(r-1)}$ and $\{\boldsymbol{\beta} \in \mathbb{R}^{J+K} : \beta_{1q_h} - \beta_{1,q_h+1} = 0\}$ is the linear restriction of interest when $t_{q_h I_{q_h}}$ is the candidate change-point to be removed.

Algorithm 1 (H) *The last K elements of $\widehat{\boldsymbol{\beta}}_{\mathcal{H}_0(K)}$, $\widehat{\beta}_{1k,\mathcal{H}_0(K)}$, $k = 1, ..., K$, are obtained solving K system of equations*

$$f_k(\widehat{\beta}_{11,\lambda}, ..., \widehat{\beta}_{1K,\lambda}) = \sum_{i=1}^{I_k} (t_{ki} - t_{k-1I_{k-1}}) \left(\sum_{j=1}^{J} D_{kji} - \sum_{j=1}^{J} m_{kji}(\widehat{\boldsymbol{\beta}}_{\mathcal{H}_0(K)}) \right) = 0,$$

$k = 1, ..., K$, *where*

$$\begin{aligned}
& m_{kji}(\widehat{\boldsymbol{\beta}}_{\mathcal{H}_0(K)}) \equiv n_{kji} \exp(\widehat{\beta}_{0j,\mathcal{H}_0(K)}) \exp(\nu_{ki}), \\
& \exp(\widehat{\beta}_{0j,\mathcal{H}_0(K)}) \equiv \frac{\sum_{r=1}^{K} \sum_{s=1}^{I_r} D_{rjs}}{\sum_{a=1}^{K} \sum_{b=1}^{I_a} n_{ajb} \exp(\nu_{ab})}, \\
& \nu_{ki} = \nu_{k-1} + \widehat{\beta}_{1k,\lambda}(t_{ki} - t_{k-1I_{k-1}}), \\
& \nu_h = \sum_{r=1}^{h} \widehat{\beta}_{1r,\mathcal{H}_0(K)}(t_{rI_r} - t_{r-1I_{r-1}}), \quad \nu_0 = 0.
\end{aligned}$$

The estimators under $\mathcal{H}_0(K - r)$, $r = 1, ..., K$, are computed in a similar way.

3 Likelihood ratio test-statistics

Before establishing the main result associated with the likelihood ratio-test statistic, we shall see a result which is valid for shortening its expression.

Proposition 1 *For each subperiod $k = 1, ..., K$,*

$$\sum_{j=1}^{J} \sum_{i=1}^{I_k} m_{kji}(\widehat{\boldsymbol{\beta}}_{\mathcal{H}_0(h)}) = \sum_{j=1}^{J} \sum_{i=1}^{I_k} D_{kji}, \; h = 0, ..., K-1.$$

Theorem 1 *The likelihood ratio test-statistic for (3) is*

$$\begin{aligned}
L(\widehat{\boldsymbol{\beta}}_{\mathcal{H}_0(h)}, \widehat{\boldsymbol{\beta}}_{\mathcal{H}_1(h)}) &= 2 \left(\ell_{\widehat{\boldsymbol{\beta}}_{\mathcal{H}_0(h)}}(\boldsymbol{D}) - \ell_{\widehat{\boldsymbol{\beta}}_{\mathcal{H}_1(h)}}(\boldsymbol{D}) \right) \\
&= 2 \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{i=1}^{I_k} D_{kji} \log \frac{m_{kji}(\widehat{\boldsymbol{\beta}}_{\mathcal{H}_0(h)})}{m_{kji}(\widehat{\boldsymbol{\beta}}_{\mathcal{H}_1(h)})},
\end{aligned}$$

and its distribution is Chi-square with one degree of freedom under $\mathcal{H}_0(h)$.

4 Simulation Study and Numerical Results

Following Tiwari et al. (2005) and taking into account the good performance of the model selection approach based on Schwartz's Bayes information criterion BIC, apart from our methodology we have analyzed by simulation the aforementioned metodology. The BIC approach selects $\boldsymbol{\beta} \in \Theta_h$, with the minimum value of

$$\mathrm{BIC}(h) = \frac{1}{I}\left(-2\ell_{\widehat{\boldsymbol{\beta}}_{\mathcal{H}_0(h)}}(\boldsymbol{D}) + h\log(I)\right),$$

where $I = \sum_{k=1}^{K} I_k$. Based on the data from the SEER, several numerical results are provided.

References

Aitchison, J., and Silvey, S. D. (1958). Maximum-Likelihood Estimation of Parameters Subject to Restraints. *Annals of Mathematical Statistics*, **29**, 813-828.

Kim, H., Fay, M., Feuer, E., and Midthune, D. (2000). Permutation tests for joinpoint regression with applications to cancer rates, *Statistics in Medicine*. **19**, 335-351.

Hall, H. I., Lee, L. M., Li J., Song, R., McKenna M. T. (2005). Describing the HIV/AIDS epidemic: using HIV case data in addition to AIDS case reporting, *Annals of Epidemiology*, **15**, 5-12.

Li, Y., Tiwari, R.C., and Zou, Z. (2008). An age-stratified model for comparing trends in cancer rates across overlapping regions. *Biometrical Journal*, **50**, 608-619.

Rosamond, W. D., Chambless, L. E., Folsom, A. R., and others (1998). Trends in the Incidence of Myocardial Infarction and in Mortality Due to Coronary Heart Disease, 1987 to 1994. *The New England Journal of Medicine*, **339**, 861-867.

Tiwari, R.C., Cronin, K.A., Davis, W., and Feuer, E.J. (2005). Bayesian model selection for join point regression with application to age-adjusted cancer rates. *Applied Statistics*. **54**, 919-939.

Weir, H.K., Thun, M.J., Hankley, B.F., Ries, L.A.G., Howe, H.L., and others (2003). Annual Report to the Nation on the Status of Cancer, 1975-2000, Featuring the Uses of Surveillance Data for Cancer Prevention and Control. *Journal of the National Cancer Institute*. **95**, 1276-1299.

Worsley, K, J. (1983). The power of likelihood ratio and cumulative sum tests for a change in a binomial probability. *Biometrika*, **70**, 455-464.

A multivariate approach for gene-sets comparison

M. Sofia Massa[1], Monica Chiogna[1], Chiara Romualdi[2]

[1] Department of Statistical Sciences, University of Padova, Via Cesare Battisti, 241, 35121, Padova, Italy. E-mail addresses: massa@stat.unipd.it, monica@stat.unipd.it

[2] Department of Biology, University of Padova, via U. Bassi 58/B, 35121, Padova, Italy. E-mail address: chiara.romualdi@unipd.it

Keywords: Gene-sets, pathway, graphical models, shrinkage covariance matrix.

1 Abstract

A microarray experiment typically provides a list of differentially expressed genes that represents the starting point of a highly difficult process of results interpretation. Biological interpretation becomes easier if differentially expressed genes show some similarity according to their functional annotation. Thus, in recent years, the interest has moved from the study of individual genes to that of groups of genes (defined by functional categories or metabolic pathways) and methods for gene set analysis have received a great attention. The aim is to identify groups of genes with moderate, but coordinated, expression changes, which should enable the understanding of cellular processes involved in the biological problem at hand. Such approaches directly score pre-defined gene sets for differential expression.
Several gene set analysis methods have been recently developed, both in the univariate and multivariate context. For a comprehensive review on existing methods see Ackermann and Strimmer (2009) and references therein.
In the multivariate perspective, Goeman and Mansmann (2008) propose Global Test, modelling differential gene expression by means of random-effects logistic regression models, while Mansmann and Meister (2005) propose ANCOVA Global Test, which is similar to Global Test but with phenotype and genes exchanged in regression models. More recently, Tsai and Chen (2009) propose a MANOVA test using a shrinkage covariance matrix estimator for the sample covariance matrix.
One of the database widely in use for the a priori definition of gene sets is the Kyoto Encyclopedia of Genes and Genomes (KEGG in the following, Ogata *et al.*, 1999), where gene products are structured into several known metabolic and regulatory pathways. A pathway is a graphical diagram of biochemical reactions involving different enzymes, where directed

and undirected edges connect few different gene products at time, according to their chemical interactions. Although KEGG pathways are usually applied to define gene sets, the approaches so far proposed do not explicitly take into consideration the dependence structure among genes implied by the topology of the pathway.

We propose to pursue the study of the behaviour of pathways in different experimental conditions within a graphical models context. This approach, whose application in the context of pathways analysis is still largely unexplored, goes in a direction which can valuably complement approaches more extensively offered by the current literature. In fact, by recording the structure of the pathway in an appropriately defined graph, we are able to keep track of the biochemical structure and reactions of the enzymes. In taking this route, the main interest is not in the detection of the structure of the pathway, because we consider it as fixed from the very beginning. In this sense, our approach differs from approaches for the analysis of differential coexpression (Gillis and Pavlidis, 2009). In other words, we are not interested in learning the structure of the pathways from the data (see Markowetz and Spang, 2007); instead, we exploit the available biological knowledge to define appropriate statistical analyses.

Within the graphical models context, data are considered as coming from Gaussian multivariate distributions with a structured concentration matrix (inverse of the covariance matrix), which reflects dependencies among variables. We present in detail two statistical tests for comparing gene sets under different experimental conditions, which naturally stem from the adopted theoretical framework and cover cases in which the number of observations is much greater than the number of variables.

The first one addresses the question of testing whether the strength of the connections among genes is altered in different experimental condition. It is likely to figure that a pathological condition does not change the structure of a pathway, but, rather, can influence the strength of the biochemical reactions. For example, a strong partial correlation among two genes in the healthy state could diminish in the disease state, or vice versa. Therefore, discovering any statistically significant difference among conditions that share the same underlying chemical structure is a crucial information. Our first test focusses, in particular, on the strength of the links among gene products and on their possible changes when considering two (or more) experimental conditions of interest.

The second test is more traditionally designed for testing for differential expression. In doing the test, we specifically employ the information about the behaviour of the partial correlations among genes and about their possible heteroschedasticity in different experimental conditions. We stress that the two tests can be performed independently one from each other.

When the number of variables exceeds the sample sizes, i.e., $n < p$, we run into the well known problem of estimation of the covariance matrices of the graphical models and of goodness of the approximation of the asymp-

totic null distribution of the test statistic. If, in addition, n is smaller than the dimension of the larger clique of the graph, we propose to obtain the covariance matrices needed for the test by running the Iterative Proportional Scaling algorithm (see Lauritzen, 1996) on starting estimates of the covariance matrices obtained via a shrinkage estimation (see Schafer and Strimmer, 2005).

The adoption of graphical methods makes it possible to decompose the overall statistical model into smaller models, with the aim of exploring in more detail small portions of the entire model. This ability naturally leads us to wish to compare portions of the pathways, with the aim of identifying subgroups of genes which appear to drive differences (deregulations) of the entire structure. The application of this idea to biological pathways is highly innovative, as it allows to look in detail to components of the pathway, opportunely defined, that can be studied separately. In fact, the expression/correlation behaviour of a large pathway could be misleading, hiding significant parts of the pathway mostly involved in the biological process under exam. With the help of the graphical models arguments, we attempt to uncover such parts.

We used a dataset recently published by Chiaretti *et al.* (2005), which characterizes gene expression signatures in acute lymphocytic leukemia (ALL) cells associated with known genotypic abnormalities in adult patients. Several distinct genetic mechanisms lead to ALL malignant transformations deriving from distinct lymphoid precursor cells that have been committed to either T-lineage or B-lineage differentiation. Chromosome translocations and molecular rearrangements are common events in B-lineage ALL and reflect distinct mechanisms of transformation. The relative frequencies of specific molecular rearrangements differ in children and adults with B-lineage ALL. The B Cell Receptor (BCR/ABL) gene rearrangement occurs in about 25% of cases in adult ALL, and much less frequently in pediatric ALL. Because these cytogenetic abnormalities reflect distinct mechanisms of transformation, molecular differences between these two types of rearrangements could help to explain why children and adults with ALL have such different outcomes following conventional therapy. Our approach identifies, as expected, the B cell receptor signaling pathway assignificantly involved in groups difference and shows that only a part of the entire pathway seems to be responsible of the different prognostic behaviour of BCR/ABL positive and negative patients. In particular, we find, in agreement with published experimental evidences, that JUN oncogene with RAS/MAPK/JNK followed by NFAT and NFKB seem to be the key regulatory elements in the comparison of BCR/ABL positive and negative patients.

References

Ackermann, M., Strimmer, K. (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10:47.

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Wang, K.S., Mandelli, F., Fo, R., Ritz, J. (2005). Gene expression profiles of B-lineage adult acute lymphocytic leukemia reveal genetic patterns that identify lineage derivation and distinct mechanisms of transformation. *Clinical Cancer Research*, 11, 7209–7219.

Gillis, J., Pavlidis, P. (2009). A methodology for the analysis of differential coexpression across the human lifespam. *BMC Bioinformatics*, 10:306.

Goeman, J.J., Mansmann, U. (2008). Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics*, 24, 537–544.

Lauritzen, S.L. (1996). *Graphical models*. Oxford: Clarendon Press.

Markowetz, F., Spang, R. (2007). Inferring cellular networks - a review. *BMC Bioinformatics*, 8:S5.

Mansmann, U., Meister, R. (2005) Testing Differential Gene Expression in Functional Groups. Goeman's Global Test versus an ANCOVA Approach. *Methods of Information in Medicine*, 44, 449–453.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27, 29–34.

Schafer, J., Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4:32.

Tsai, C.A., Chen, J.J. (2009). Multivariate analysis of variance test for gene set analysis. *Bioinformatics*, 25, 897–903.

Identifiability of causal effects in randomized experiments with noncompliance, nonignorable missing data, and binary outcomes

Andrea Mercatanti[1]

[1] Bank of Italy, Economic and Financial Statistics Department; Via Nazionale 91, 00184 Roma, Italy. e-mail: andrea.mercatanti@bancaditalia.it

Abstract: The theoretical model of a randomized experiment with noncompliance has been widely adopted in identifying and estimating causal effects in economic and social sciences. In these cases, information is generally collected from surveys, wherein missing data is a common problem; therefore standard methods for complete data cannot be immediately used. Moreover, possible biases exist because the respondents are often systematically different from the nonrespondents; of particular concern, these biases are difficult to eliminate since the precise reason for nonresponse are usually not known. However, the strand of literature dealing with missing data is typically restricted to the field of biostatistics, where nonignorable conditions for the missing data mechanism have been proposed for situations of missingness only in the outcome. There are various situations in economic and social sciences, where data may not only be missing in the outcome, and where usual ignorability conditions for the missing data mechanism are considered too restrictive. In this context, I propose a way to obtain identifiable models by the introduction of suitable parameter restrictions under nonignorability conditions for the missing data mechanism.

Keywords: Potential Outcomes Approach; Missing Data; Noncompliance.

1 Notation

Adopting the standard notation in the literature on randomized experiments with noncompliance and missing data, we indicate with: D the binary treatment; Z the binary assignment to treatment; Y the binary outcome; U the compliance status, $U = a$ (always-takers), n (never-takers), c (compliers); $\mathbf{R}$ the three-dimensional vector (R_y, R_d, R_z), where R_v is the binary missing data indicator for $v = y, d, z$: $R_v = 1$ if v is observed, 0 otherwise; D^*, Z^*, and Y^* the observable quantities:

- $D^* = D$ if $R = 1$, $D^* = *$ if $R = 0$
- $Z^* = Z$ if $R = 1$, $Z^* = *$ if $R = 0$

- $Y^* = Y$ if $R = 1$, $Y^* = *$ if $R = 0$

2 Assumptions

We maintain the standard assumptions to identify causal effects in randomized experiments with noncompliance (Angrist at al., 1996) throughout the paper; these are: the Stable Unit Treatment Value Assumption; the randomization of Z; the nonzero effect of Z on D; the absence of defiers; the exclusion restriction for noncompliers.

3 Identifiability

Under the previous set of assumptions and maintaining the parameter space of the model for $(Y,\, D,\, Z)$ separated from the parameter space of the missing data mechanism, the general model for the observable data can be written as:

$$P(\mathbf{R},\, Y^*,\, D^*,\, Z^*;\, \pi,\, \omega,\, \theta,\, \alpha) =$$

$$= \sum_{Y,\, Z,\, U} P(\mathbf{R},\, Y,\, U,\; Z;\, \pi,\, \omega,\, \theta,\, \alpha) \cdot I[P(Y \,|\, Y^*) > 0] \cdot I[P(Z \,|\, Z^*) > 0] \cdot$$

$$\cdot I[P(D(Z) = D^* | U,\, Z^*, D^*) > 0] =$$

$$= \sum_{Y,\, Z,\, U} P(\mathbf{R}|\, Y,\, U,\, Z; \alpha) \cdot P(Y,\, U,\, Z;\, \pi,\, \omega,\, \theta) \cdot I[P(Y \,|\, Y^*) > 0] \cdot$$

$$\cdot I[P(Z \,|\, Z^*) > 0] \cdot I[P(D(Z) = D^* | U,\, Z^*, D^*) > 0]. \qquad (1)$$

We impose the three marginal missing data mechanisms $P(R_v|\, Y,\, U,\, Z)$, $v = Y,\, D,\, Z$, to be pairwise conditionally independent given $(Y,\, U,\, Z)$. The restriction allows us to specify three binary independent models, $P(R_v|\, Y,\, U,\, Z)$, instead of a joint model $P(\mathbf{R}|\, Y,\, U,\, Z)$ whose domain would be the entire set of values for $\mathbf{R} = (R_Y,\, R_D,\, R_Z) : \{(1,1,1)\,,\, (1,1,0)\,,\, (1,0,0)\,.....\}$. Under this assumption, a general saturated model for $P(\mathbf{R},\, Y^*,\, D^*,\, Z^*;\, \pi,\, \omega,\, \theta,\, \alpha)$ can be proposed by specifying the model for $(Y,\, U,\, Z)$ like in Imbens and Rubin (1997):

$$P(Y,\, U,\, Z;\; \pi, \omega, \theta) = \pi^Z \,(1-\pi)^{1-Z}\, \omega_a^{I(U=a)}\, \omega_n^{I(U=n)}$$

$$(1-\omega_a-\omega_n)^{I(U=c)}\, \theta_a^{Y\, I(U=a)}\, (1-\theta_a)^{(1-Y)\, I(U=a)}\, \theta_n^{Y\, I(U=n)}\, (1-\theta_n)^{(1-Y)\, I(U=n)}$$

$$\theta_{c1}^{Y\,I(U=c)\,Z}\,(1-\theta_{c1})^{(1-Y)\,I(U=c)\,Z}\,\theta_{c0}^{Y\,I(U=c)\,(1-Z)}(1-\theta_{c0})^{(1-Y)\,I(U=c)\,(1-Z)}, \tag{2}$$

and by adopting, for each marginal model $P(R_v|\,Y,\,U,\,Z)$, the missing data mechanism recently proposed by Small and Cheng (2008), here extended to allow for interactions between the outcome and the compliance status:

$$P(\mathbf{R}|\,Y,\,U,\,Z;\,\alpha) = P(R_Y = 1|\,Y,\,U,\,Z)\cdot P(R_D = 1|\,Y,\,U,\,Z)\cdot P(R_Z = 1|\,Y,\,U,\,Z), \tag{3}$$

where

$$logit[P(R_v = 1|\,Y,\,U,\,Z)] = \alpha_{0v} + \alpha_{1v}\,Y + \alpha_{2v}\,I(U = a) + \alpha_{3v}\,I(U = n) +$$

$$+\alpha_{4v}\,I(U = a,\,Z = 1) + \alpha_{5v}\,I(U = n,\,Z = 1) + \alpha_{6v}\,I(U = c,\,Z = 1) +$$

$$+\alpha_{7v}\,Y\cdot I(U = a) + \alpha_{8}Y\cdot\,I(U = n) +$$

$$+\alpha_{9v}\,Y\cdot I(U = a,\,Z = 1) + \alpha_{10v}\,Y\cdot I(U = n,\,Z = 1) + \alpha_{11v}\,Y\cdot I(U = c,\,Z = 1), \tag{4}$$

for $v = Y,\,D,\,Z$.

A simple way to obtain parameter identification stems from the analysis of the contingency table for the observable data. In general the number of free parameters for regular models has to be at most the number of cells of the corresponding contingency tables minus one. In the current situation, the resulting contingency table for the observable data allows at most 26 free parameters. The particular form of (2), with 7 parameters, stems from the presence of two mixtures of distributions, while (3) is specified as a sequence of three independent logistic regression each presenting 12 parameters. The saturated model for $P(\mathbf{R},\,Y^*,\,D^*,\,Z^*;\,\pi,\,\omega,\,\theta,\,\alpha)$ has in total $7+(12\times3) = 43$ parameters, consequently at least $43 - 26 = 17$ constraints have to be imposed.

The pairwise conditional independence of R_y, R_d and R_z given $(Y,\,U,\,Z)$ can be relaxed at the cost of introducing a binary observable covariate X related to $(Y,\,U,\,Z)$, so that the missing data mechanism can be modelized as a multinomial logit with eight categories: $\mathbf{R} = (R_Y,\,R_D,\,R_Z)$: $\{(1,1,1)\,,\,(1,1,0)\,,\,(1,0,0)\,.....\}$. The resulting contingency table for the observables now allows at most 53 free parameters for regular models. The component $P(Y,\,U,\,Z,\,X)$ of the saturated model for the observables can now be expanded up to 14 parameters:

$$P(Y,\,U,\,Z,\,X;\;\pi,\omega,\theta) = \pi^{Z}\,(1-\pi)^{1-Z}\,\omega_{1a}^{I(X=1,\,U=a)}\,\omega_{0a}^{I(\,X=0,\,U=a)}\,\omega_{1n}^{I(\,X=1,\,U=n)}$$

$$\omega_{0n}^{I(X=0,\,U=n)}\,\omega_{1c}^{I(X=1,\,U=c)}\,(1-\omega_{1a}-\omega_{0a}-\omega_{1n}-\omega_{0n}-\omega_{1c})^{I(X=0,\,U=c)}\,\theta_{1a}^{Y\,I(X=1,\,U=a)}$$

$$\theta_{0a}^{Y\,I(X=0,\,U=a)}\,(1-\theta_{1a})^{(1-Y)\,I(X=1,\,U=a)}\,(1-\theta_{0a})^{(1-Y)\,I(X=0,\,U=a)}\theta_{1n}^{Y\,I(X=1,\,U=n)}$$

$$\theta_{0n}^{Y\,I(X=0,\,U=n)}\,(1-\theta_{1n})^{(1-Y)\,I(\,X=1,\,U=n)}\,(1-\theta_{0n})^{(1-Y)\,I(X=0,\,U=n)}\theta_{1c1}^{Y\,I(X=1,\,U=c)\,Z}$$

$$\theta_{0c1}^{Y\,I(X=0,\,U=c)\,Z}\,(1-\theta_{1c1})^{(1-Y)\,I(X=1,\,U=c)\,Z}\,(1-\theta_{0c1})^{(1-Y)\,I(X=0,\,U=c)\,Z}\theta_{1c0}^{Y\,I(X=1,\,U=c)\,(1-Z)}$$

$$\theta_{0c0}^{Y\,I(X=0,\,U=c)\,(1-Z)}\,(1-\theta_{1c0})^{(1-Y)\,I(X=1,\,U=c)\,(1-Z)}\,(1-\theta_{0c0})^{(1-Y)\,I(X=0,\,U=c)\,(1-Z)}.$$

Under the assumption that X does not enter the missing data mechanisms each parameter vector $\alpha_{(R_Y,\,R_D,\,R_Z)}$ of the multinomial logit for $(R_Y,\,R_D,\,R_Z)$ contains 12 terms, like in (4). The saturated model has now $14+(12\times 7)=98$ so that at least $98-53=45$ constraints have to be imposed on $P(Y,\,U,\,Z;\;\pi,\omega,\theta)\cdot P(\mathbf{R}|\,Y,\,U,\,Z;\,\alpha)$.

4 Conclusions

The proposed method allows to specify alternative identifiable models by the introduction of a minimum number of parameter restrictions on the saturated model. The restrictions have to be imposed in order to make equal the maximum number of parameters to the number of cells of the contingency table for the observables minus one.

Acknowledgments: Special Thanks to Guido W. Imbens for useful comments and suggestions

References

Angrist J.D., G.W. Imbens, D.B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, **91**.

Chen H., Z. Geng, X.H. Zhou (2008). Identifiability and estimation of causal effects in randomized trials with noncompliance and completely nonignorable missing data. *Biometrics, Published Online: 28 Aug 2008.*.

Imbens G.W., D.B. Rubin (1997). Bayesian inference for causal effects in randomized experiments with noncompliance *The Annals of Statistics*, **25**.

Sigmoid Models Utilized in Optimization of Gas Transportation Networks

Radoslava Mirkov [1], Herwig Friedl [2], Hernan Leövey [1], Werner Römisch [1], Isabel Wegner-Specht [1]

[1] Humboldt Universität zu Berlin, Department of Mathematics, Unter den Linden 6, 10999 Berlin, Germany, mirkov@math.hu-berlin.de
[2] Graz University of Technology, Institute of Statistics, Münzgrabenstraße 11, 8010 Graz, Austria

Abstract: The flow of natural gas within a gas transmission network is studied with the aim to optimize such networks. The analysis of real data provides a deeper insight into the behavior of gas in- and outflow. A sigmoid regression model is chosen to describe dependence between the maximal daily gas flow and the temperature on network exits.

Keywords: Sigmoid Regression; Gas Transport; Optimization.

1 Introduction

Transportation and supply of natural gas is an important topic, and we study the flow of gas transported in networks in the past in order to support the optimization of such networks and thus improve the supply of gas. To do so we fit a nonlinear regression model and analyze the properties of the gas flow through the pipelines in dependence of the temperature.
Data is obtained from measuring stations within the German pipeline network, and contains hourly gas flow for the period of the last five years. Mean daily temperatures are also provided. We study the dependence between gas consumption and air temperature on all exits along the pipelines. Since we want to maximize the transportation capacity through the pipelines, we concentrate on the daily maximum flows y_i, $i = 1, \ldots, n$, at each exit, for every exit in the network.

2 Sigmoid Regression

In what follows, we concentrate on the data observed at one specific station. Based on the Cooperation Agreement (2008) between gas companies, we choose the following sigmoidal growth model to describe the dependence of gas consumption on temperature:

$$y_i = \mu \times S(\theta|t_i) + \varepsilon_i, \tag{1}$$

where μ denotes the overall mean of all maximal daily gas flows at one specific measuring station, t_i stands for the weighted four-day-mean temperature with weights $(0.5333, 0.2667, 0.1333, 0.0667)$, the sigmoid function S with parameter $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ is given by

$$S(\theta|t_i) = \theta_4 + \frac{\theta_1 - \theta_4}{1 + \left(\frac{\theta_2}{t_i - 40}\right)^{\theta_3}}, \tag{2}$$

and $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ are error terms.

The choice of this model is based on physical properties of gas, see e.g. Cerbe (2008). The four parameters in model (2) have the following meaning: θ_1 and θ_4 are the upper and lower asymptotes, and the other two parameters describe the shape of the decrease of the (logistic like) curve. More precisely, θ_2 is the inflection point of the curve, and θ_3 is proportional to the slope at $t = \theta_2$ (cf. Ritz and Streibig, 2008).

The starting values for the iteration necessary to calculate the maximum likelihood (ML) estimates, as provided in the Cooperation Agreement (2008), are given in the Table 1. Alternatively, according to Seber and Wild (2003), we obtain the crude initial estimates of θ_1 and θ_4 from the scatterplot, while θ_2 and θ_3 can be obtained using the linearization

$$y^* = \theta_3 \log(-\theta_2) - \theta_3 \log(-t + 40).$$

Note that the starting values, which have been estimated from data are much closer to the fitted parameters than the values given in the Cooperation Agreement (2008), whereas both sets of starting values yield the same model parameters. The model specified by (1) and (2) is fitted in R using the function `nls()`. We refer to Table 2 for the estimated parameters in the fitted model.

TABLE 1. Initial values of the parameters in the sigmoid model.

Method	θ_1	θ_2	θ_3	θ_4
Agreement (A)	2.5086	−34.7213	5.8164	0.1208
Estimated (E)	1.9111	−32.0284	6.0055	0.4375

TABLE 2. ML estimates (standard errors) under the sigmoid model.

θ_1	θ_2	θ_3	θ_4
2.0330	−32.647	6.6644	0.4468
(0.0316)	(0.2144)	(0.2185)	(0.0093)

In Figure 1(left) we show the model fitted to data describing the typical gas outflow for public utilities, as well as the curves corresponding to both

sets of starting values (A and E) given in Table 1. This figure suggests that the sigmoid model (M) reproduces the gross characteristics of the gas flow well, though it obviously underestimates the mean responses for low temperatures.

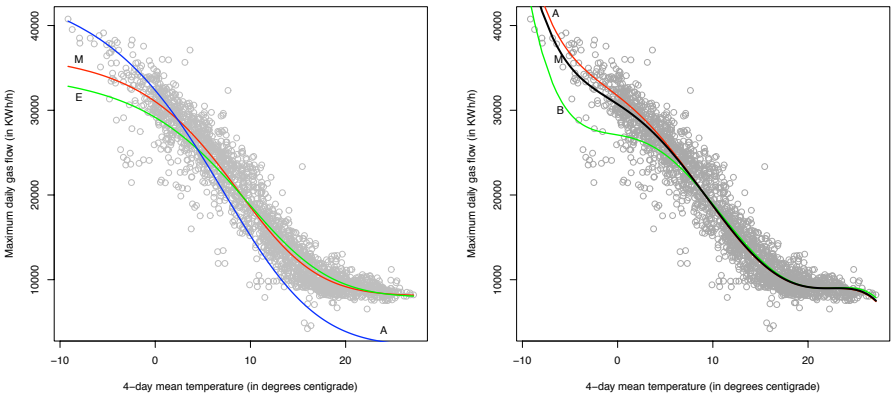


FIGURE 1. Fitted sigmoid (left) and FlexMix model (right).

3 Mixture of Polynomial Regression Models

The nuances missed by the sigmoid model motivates the use of the FlexMix approach introduced in Leisch (2004), which offers a framework for the flexible fitting of finite mixtures of regression models in the R environment. Since this framework mostly covers generalized linear models, we approximate the nonlinear sigmoid function S in (2) by a polynomial of the fifth degree, and fit a two-component mixture model

$$\text{Class A: } y_i = \sum_{j=0}^{5} \beta_j^A t_i^j + \varepsilon_i, \qquad \text{Class B: } y_i = \sum_{j=0}^{5} \beta_j^B t_i^j + \varepsilon_i. \tag{3}$$

The model given by (3) is fitted in R using the function `flexmix()`, the parameters of the fitted model are given in Table 3. Table 4 shows the estimated priors, the number of observations assigned to the clusters, the number of observations with posterior probabilities larger than 0.0001 and the ratio of the latter two numbers. The ratio of the second component is approximately 0.1, indicating the overlap of the classes for the large proportion of data. Figure 1(right) shows both conditional models (A and

TABLE 3. Estimates (standard errors) under the FlexMix model.

Class	β_0	β_1	β_2	β_3	β_4	β_5
A	31841	−901.85	−42.44	−3.165	0.402	−0.008
	(113.12)	(25.709)	(4.334)	(0.382)	(0.009)	
B	26714	−895.16	−7.197	−13.867	0.957	−0.017
	(622.74)	(162.46)	(32.046)	(2.489)	(0.059)	

TABLE 4. Mixed regression properties.

Class	Prior Prob.	Cluster Size	Posterior>0	Ratio	σ
A	0.779	1814	1972	0.9199	1506
B	0.221	191	2005	0.0953	3624

B) as also the marginal model (M) obtained from the FlexMix procedure. The fit for low temperatures differs clearly for both model components, and describes the gas flow in a more appropriate way.

4 Conclusions

Preliminary results show that a simple sigmoid model enables a good starting approach to the observed problem. The shape of the sigmoid function we used to model the dependence of the maximum gas flow from temperature is suitable, but there is room for improvement. We suggest the flexible fitting of mixture models. Since the approximation by polynomials is unstable, the FlexMix approach would offer even better results, if one could generalize this method to mixtures of sigmoid models.

References

Cerbe, G. (2008). *Grundlagen der Gastechnik*. Hanser Verlag.

Cooperation Agreement (2008). *Vereinbarung über die Kooperation gemäß* §20 *Absatz* 1 *b) EnWG zwischen den Betreibern von in Deutschland gelegenen Gasversorgungsnetzen*. BMJ Deutschland.

Leisch, F. (2004). FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. *Journal of Statistical Software*, **11**, 1-18.

Ritz, C., and Streibig, J.C. (2008). *Nonlinear Regression with R*. Springer.

Seber, G.A.F., and Wild, C.J. (2003). *Nonlinear Regression*. New Jersey: John Wiley & Sons.

A statistical model for hospital admissions caused by seasonal diseases

D. Moriña[1,2], P. Puig[2], A. Vilella[3], A. Trilla[4]

[1] Centre Tecnològic de Nutrició i Salut, Reus
[2] Servei d'Estadística, Universitat Autònoma de Barcelona
[3] Servei de Medicina Preventiva, Hospital Clínic i Provincial de Barcelona
[4] Unitat d'Avaluació, Suport i Prevenció, Hospital Clínic i Provincial de Barcelona

Abstract: We present an example of discrete time series analysis related with the number of hospital admissions per week due to causes attributable to flu. The constructed models are based on Poisson innovations.

Keywords: Discrete time series; INAR(1) models; Seasonality; Hospital admissions model.

1 Introduction

Unfortunately, the decision to hospitalize or not a patient that reaches a hospital emergency service is not taken only in accordance with its gravity, but it has to do with the number of beds available. It is therefore crucial to make reliable predictions of the number of beds that will be available in the future depending on what happened in the past. This paper attempts to answer this problem using discrete time series (see McKenzie, 2000). Some diseases, such as flu, occur more frequently at certain times of the year. This seasonal behavior is not covered by discrete time series classical models INAR(1), defined by

$$X_t = p \circ X_{t-1} + W_t, \tag{1}$$

where p is a fixed parameter, $0 < p < 1$ and W_t is assumed to follow a Poisson distribution with mean λ. X_{t-1} and W_t are assumed to be independent at any time t. A good review of these models can be found in Jung and Tremayne (2006).
The $\circ$ operator, called *binomial thinning* is defined as follows,

$$p \circ X_{t-1} = \sum_{i=1}^{X_{t-1}} Y_i, \tag{2}$$

where Y_i follow a Bernoulli distribution with probability of success equal to p. Therefore, $p \circ X_{t-1}$ is a binomial random variable where the number

of experiments is X_{t-1}.
Usually, the parameter p can be interpreted as the proportion of observations counted on time $t-1$ that remains on time t. Depending on the context, it can be understood as a survival tax. On the other hand, W_t can be interpreted as the innovations (or repositions) produced at time t. A characterization of count distributions based on p-thinning properties can be found in Puig and Valero (2007). These distributions, where Poisson is one of them, can be used as appropriate innovations.
The theoretical model defined in the next section will be applied to a set of data concerning the number of arrivals per week to the emergency service of the *Hospital Clínic i Provincial de Barcelona* due to causes attributable to flu. Data was collected between January of 2004 and December of 2008. Based on this model, our goal is to study the behavior of this phenomenon and to make predictions.
In our case, we can consider the parameter p as the proportion of patients who went to emergency service at time $t-1$ and had a relapse the next week, and W_t can be interpreted as the new patients who arrive to the emergency service at time t.

2 Model definition

To include the seasonal impact in the model, we propose the following variation to the INAR(1) model defined in (1):

$$X_t = p \circ X_{t-1} + W_t(\lambda_j). \tag{3}$$

Here the parameter p is as before but now the innovations W_t are supposed to follow several Poisson distributions with different means $\lambda_1, \ldots, \lambda_n$, where n catches the estimated periodicity. Still, X_{t-1} and W_t are supposed to be independent. This model is similar to that of Zhu and Joe (2006). In fact we are using the same time as a covariate.
We have assumed a seasonality or periodicity of 12 months in our data after an exploratory approach based on lineal regression analysis. To do that, we have used a simple one-wave trigonometric model defined by

$$Z_t = \alpha + \beta \sin\left(\frac{2\pi t}{T}\right) + \gamma \cos\left(\frac{2\pi t}{T}\right), \tag{4}$$

where T is the period. We have found that the best fit corresponds to a choice of $T = 12$. As we can see in Figure 1, the regression model is far from fitting the real data, although it captures its seasonality. Therefore, there will be a total of 13 parameters to be estimated, $p, \lambda_1, \ldots, \lambda_{12}$.

3 Parameter estimation

We have estimated the parameters of the model by using conditioned maximum likelihood, taking into account the fact that we can write the likeli-

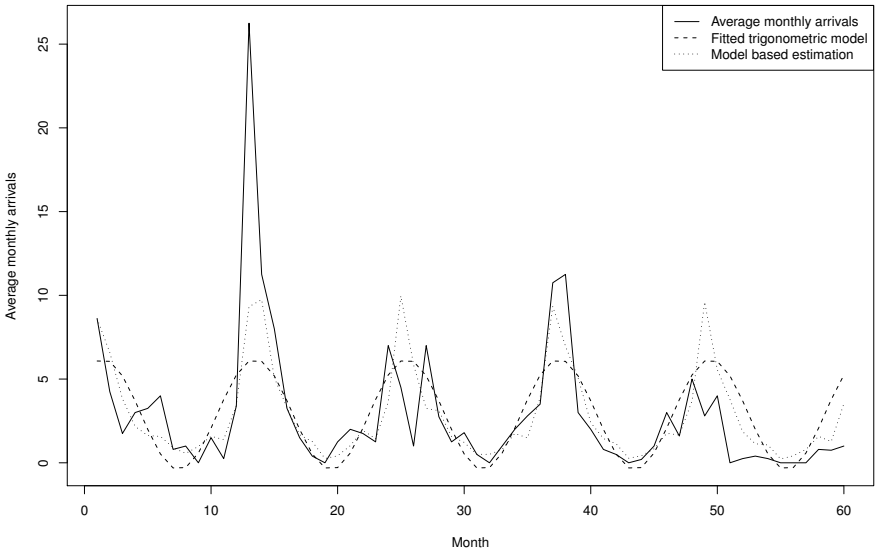


FIGURE 1. Observed data, fitted trigonometric model and the model based mean prediction

hood function as

$$L(X;\theta) = f(x_1,\ldots,x_n) = f(x_1)\cdot f(x_2 \mid x_1)\cdots f(x_n \mid x_{n-1}), \qquad (5)$$

because we are considering only dependencies of order 1. Therefore, we have to compute the probability function of the sum of a binomial variable with parameters (X_{i-1}, p) and an independent Poisson variable with mean $\lambda_1,\ldots,\lambda_{12}$ depending on the month. To do that we have used the following well known result:

Proposition. *Let U be a binomial random variable with parameters (m,p), and let V be a Poisson random variable with mean λ. Suppose that U and V are independent. Then, the probability function of $Z = U + V$ can be written in the form*

$$P(Z=s) = e^{-\lambda}\cdot \sum_{k=0}^{min(m,s)} \frac{m!\lambda^{s-k}}{(m-k)!k!(s-k)!}\cdot p^k\cdot(1-p)^{m-k} \qquad (6)$$

Therefore, the conditioned likelihood function can be expressed as

$$L(X;\lambda_1,\ldots,\lambda_{12},p) = \prod_{i=2}^{n} P(x_{i-1},x_i), \qquad (7)$$

where

$$P(x_{i-1}, x_i) = e^{-\lambda_j} \cdot \sum_{k=0}^{min(x_{i-1}, x_i)} \frac{x_{i-1}! \lambda_r^{x_i - k}}{(x_{i-1} - k)! k! (x_i - k)!} \cdot p^k \cdot (1-p)^{x_{i-1} - k} \quad (8)$$

Here λ_j, $j = 1, 2, ..., 12$ is the population mean of the innovation corresponding to the month of the observation x_i.
The maximization of (7) has been done with a program developed in R that is available from the authors on request. The obtained estimates with their standard errors are showed in table 1.

TABLE 1. Maximum likelihood estimates.

Parameter	p	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}	λ_{11}	λ_{12}
Estimate	0.18	8.71	5.05	3.07	1.86	1.09	0.99	0.19	0.42	0.80	1.58	1.12	3.40
St. Error	0.03	0.71	0.58	0.43	0.32	0.24	0.24	0.1	0.14	0.20	0.29	0.25	0.41

Note how the largest value of the estimated mean of the innovations (8.71) corresponds to January. In fact, we have also observed an outlier in January 2005. Moreover, according to the estimated value of p, we can conclude that a 18% of the patients come back to the emergency service later on. Figure 1 also shows an estimation of the average monthly arrivals conditioned to the preceding observation, calculated using the recurrence relation,

$$\mu_t = p \cdot X_{t-1} + \lambda_t, \quad (9)$$

where $E(X_t) = \mu_t$ is the mean of the number of arrivals at time t. The profile of this plot agrees with the observed data.

4 Model validation and forecasting

We have used the recurrent relation

$$\hat{\mu}_t = \hat{p} \cdot X_{t-1} + \hat{\lambda}_i, i = 1, \ldots, 12 \quad (10)$$

to validate how the model fits the real data. We can see that Figure 2 shows that the model estimated values agrees with the observed data for the period January-December 2008. We can also see the seasonal impact in a prediction for the next year.
One can iterate expression (10) in order to obtain a prediction of future values. If X_n is the last known value, we have

$$\tilde{X}_{n+k} = \hat{p}^k \cdot X_n + \sum_{i=1}^{k} \hat{p}^{k-i} \hat{\lambda}_i, \quad (11)$$

and a confidence interval for the prediction can be obtained using the *delta method* to compute the variances, considering each $\tilde{X}_j$ as a function of the estimates of the parameters $p, \lambda_1, \ldots, \lambda_{12}$.

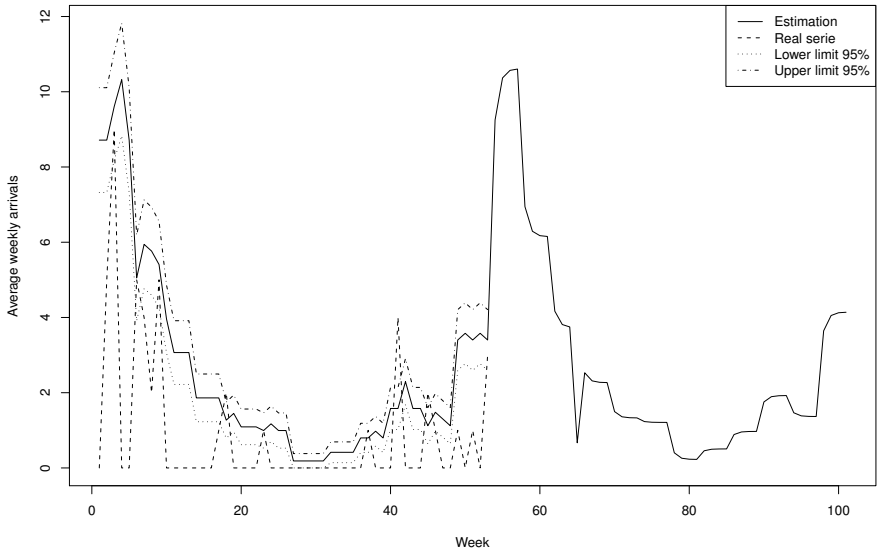


FIGURE 2. Estimation, real data and confidence interval for 2008 and prediction

Acknowledgments: We would like to thank José Ríos, from Hospital Clínic i Provincial de Barcelona, who provided us the data set that motivated this paper. This research has been partially supported by grant MTM2009-10893 from the Ministry of Education of Spain.

References

Jung, R.C. and Tremayne, A.R. (2006). Binomial thinning models for integer time series. *Statistical Modelling*, **6**, 81-96.

McKenzie, E. (2000). Discrete variate time series. *Handbook of Statistics*, **21**, 573-606.

Puig, P. and Valero, J. (2007). Characterization of count data distributions involving additivity and binomial subsampling. *Bernoulli*, **13**, 2, 544-555.

Zhu, R. and Joe, H. (2006). Modelling count data time series with Markov processes based on binomial thinning. *Journal of time series analysis*, **27**, 5, 725-738.

LASSO regression via smooth L_1-norm approximation

Vito M.R. Muggeo[1]

[1] Dip. Scienze Statistiche e Matematiche 'Vianelli', Università di Palermo

Abstract: This paper discusses estimation of regression model with LASSO penalty when the L_1-norm is replaced with its parametric smooth approximation. The resulting parameter estimators are more manageable than those from standard LASSO, standard errors are easy computed via a sandwich formula, and the model degrees of freedom may be computed straightforwardly. Moreover the resulting objective function may be minimized using usual optimization algorithms for regular models, for instance Newton-Raphson or iterative least squares.

Keywords: LASSO; L_1-norm; smooth models; least squares.

1 Introduction

The *Least Absolute Shrinkage and Selection Operator*, LASSO, was introduced by Tibshirani (1996) as a device to obtain sparse solution in linear regression models with a large number of covariates. 'Sparse solution' means that in the final solution some of the estimated β_js regression coefficients are automatically set to zero by the procedure, so allowing (continuous) variable selection and parameter estimation *simultaneously*. The key for this crucial and important feature is the L_1-norm penalty $\lambda \sum_j |\beta_j|$ which controls the sparseness of the solution via the tuning parameter $\lambda \geq 0$: at $\lambda = 0$ the solution corresponds to the OLS estimates (if these exists), and as λ increases more estimates are set to zero. For observed responses y_i and covariate vector $x_i \in R^p$, the L_1 penalized loss function for the regression model $\mu_i = x_i^T \beta$ may be written $\sum_i (y_i - x_i^T \beta)^2 + \lambda \sum_j |\beta_j|$.
Unfortunately LASSO does not come without concerns. Standard errors for the parameter estimates are not easily obtained in the L_1-penalized framework and some approximations have been discussed. Tibshirani (1996) proposed to use the ridge-type approximation $\sum_j |\beta_j| \approx \sum_j \beta_j^2 / |\tilde{\beta}_j|$ to justify a sandwich-type formula, where the tilde means an approximate known value. Osborne *et al.* (2000) showed that this approach breaks down as it provides $\mathrm{SE}(\hat{\beta}_j) = 0$ when $\hat{\beta}_j = 0$, which is inappropriate as also zero estimates would be associated with some degree of uncertainty; Osborne and co-workers derived a formula for covariance matrix which ensures positive standard errors for all coefficient estimates. A bootstrap approach was also

proposed (Tibshirani, 1996), but it fails to be consistent (Kyung et al., 2010) and moreover it may be troublesome in some contexts, especially in large dataset and/or complex models. Regardless of the approach used to compute standard errors, the sampling distribution of the regression parameter estimator with L_1 penalty is not multivariate Normal when at least one true coefficient is zero. Typically, the density of the sampling distribution with null parameter has positive probability mass at zero (Knight and Fu, 2000: Pötscher and Leeb, 2009; Kyung *et al.*, 2010). Non-normality of sampling distribution causes the standard errors to be somewhat useless for inference purpose, namely to build confidence intervals and to carry out hypothesis testing. Finally, from a practical viewpoint, although LASSO algorithms are well established in linear or even generalized linear models (Efron et al., 2004), their use may result not straightforward for more complex contexts, for instance models with heteroscedastic/autocorrelated errors and/or random effects.
In this paper we describe a smooth parametric approximation of the L_1-norm which allows to perform LASSO regression via iterative least squares and to get valid standard errors for all the parameter estimates.

2 Methods

We replace the usual L_1 penalty by its smooth and parametric approximation, which we call 'quasi-L_1', briefly qL_1; we refer the resulting qL_1 penalized regression as 'quasi-LASSO'. To begin with, we approximate the absolute value function,

$$|\beta| \approx Q(\beta) = -\beta + \frac{(\beta + c)^2}{2c} I(|\beta| \leq c) + 2\beta I(\beta > c), \tag{1}$$

where $I(\cdot)$ is the usual indicator function and c is a small known constant regulating the width of the bend around zero. Unlike $|\beta|$, $Q(\beta)$ is two-times differentiable at the origin for $c > 0$, and as $c \to 0$, $Q(\beta) \to |\beta|$. Using (1) the qL_1 penalty for a parameter vector is naturally given by $\sum_j |\beta_j| \approx \sum_j Q(\beta_j)$.
Figure 1 illustrates the LASSO and the quasi-LASSO penalties in the plane (β_1, β_2) for three different values of c; the smooth arc (for $c > 0$) replacing the kink in the neighbour of $\beta_j = 0$ guarantees differentiability of the objective function. The minimand loss function with the quasi-LASSO may be written as

$$\mathcal{L}_c = \sum_i (y_i - x_i^T \beta)^2 + \lambda \sum_j Q(\beta_j). \tag{2}$$

The main appealing of the function $Q(\cdot)$, and thus of qL_1, is that it admits first and second derivatives, and as a consequence, there exist the score and the hessian of the objective function (2). It can be shown that the

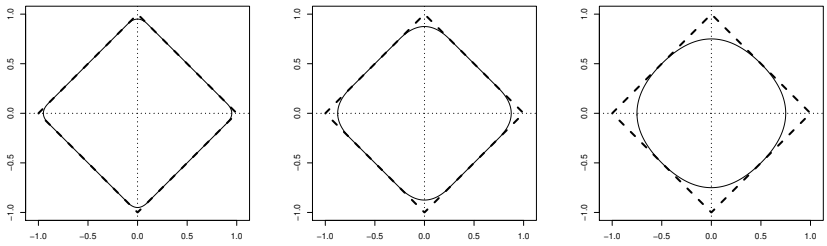


FIGURE 1. Contour plots of the LASSO (dashed lines) and the approximate LASSO (continuous lines) penalties for two coefficients and three different values of c.

Newton-Raphson step is equivalent to the least squares computation

$$\hat{\beta} = (X^T X + \lambda \widetilde{V})^{-1} X^T \tilde{z}^*, \tag{3}$$

with

$$\begin{aligned} z^* &= X^- X^T y + \lambda X^- (I_p 1_p - c\widetilde{V} 1_p - \widetilde{W} 1_p) \\ &= y + \lambda X^- (I_p 1_p - c\widetilde{V} 1_p - \widetilde{W} 1_p), \end{aligned} \tag{4}$$

where X is the usual $n \times p$ design matrix, X^- is the $n \times p$ Moore-Penrose generalized inverse of X^T, and y is response vector; moreover I_p is the $p \times p$ identity matrix and 1_p is a p-vector of ones, while $\widetilde{V} = c^{-1}\text{diag}(I(|\tilde{\beta}_j| \leq c))$ and $\widetilde{W} = \text{diag}(2I(\tilde{\beta}_j > c))$ are p-dimensional square matrices. Tilde values mean approximate values: at the first step, the initial guesses for the beta parameters may be obtained by standard OLS or even trivial starting values $(0, 0, \ldots, 0)^T$ may be employed; some empirical experience suggest that starting values are a minor issue and convergence is achieved in a few iterations. However whether several models have to be fitted (for instance to search for the optimal value of the tuning parameter) appropriate starting values may speed convergence overall. Note when additional covariates have to be included without penalizing corresponding coefficients (for instance the intercept), it suffices to include them in the design matrix X and to add zeroes to the main diagonals of the matrices V and W, such that these become blockdiag$(V, 0, \ldots, 0)$ and blockdiag$(W, 0, \ldots, 0)$.

Unlike the pure LASSO, owing to the parameterization of $Q(\beta)$ some estimates will never be exactly zero. However as the parameter c measures the amount of approximation at the kinks $\beta_j = 0$, in a 'model selection' context it is reasonable to consider zero the estimates fulfilling $|\hat{\beta}_j| \leq c$.

Parameter estimates depend on $c > 0$ and $\lambda \geq 0$ jointly. At $\lambda = 0$ the algorithm yields the OLS estimates regardless the value of c, while when $\lambda \neq 0$ the value of c is not trivial: if each $|\hat{\beta}_j| < c$, it is easy to see that $V = c^{-1} I_p$

and $W = 0$, hence $\tilde{z}^* = y$ and solution (3) reduces to the ridge penalty solution with tuning parameter λ/c. As sketched in the Introduction, it is possible to get asymptotic standard errors for the overall parameter vector estimate within the quasi-LASSO framework. The resulting sandwich formula is

$$\text{cov}(\hat{\beta}) = E[H]^{-1}\text{cov}(U)E[H]^{-T} \tag{5}$$

where H is the hessian matrix, $\text{cov}(U)$ is the covariance matrix of the gradient U, and '$-T$' means the transpose of the inverse. To apply the sandwich formula in practice, $E[H]$ is estimated by the hessian H evaluated at $\hat{\beta}$, and $\text{cov}(U) = \hat{\sigma}^2 X^T X$. The sandwich formula does not constrain to zero the standard errors for the (quasi) zero estimates and it works for any λ and c; moreover also note that the sampling distribution of the qL_1 estimates is Normal, so formula (5) may be applied to build valid confidence intervals even when the true parameter is zero.

3 Application and Simulation

We use the quasi-LASSO penalty to the well-known Prostate Cancer dataset analyzed by Tibshirani (1996). There are $n = 97$ subjects, $p = 8$ covariates (see Table 1) and the response variable is the log of prostrate specific antigen. Figure 2 illustrates the GCV and the BIC scores for the selection of the tuning parameter λ and the threshold c; to compute GCV and BIC we use the trace of the hat matrix XHX^T which is defined for the qL_1 penalty as the hessian H exists. Both GCV and BIC favour the smallest value of c which suggests sparsity in the solution; given $c = 0.1$ the BIC selects a somewhat more parsimonious model (i.e. larger λ). This is not surprising as it is well-known the BIC tends to select simple models (Zou et al., 2007).

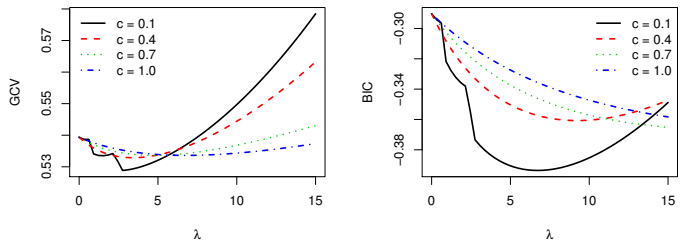


FIGURE 2. GCV and BIC scores with respect to λ for 4 values of c for the qL_1 regression for the Prostate Cancer dataset.

Table 1 shows the parameter estimates for the qL_1 regression. As reported in literature (e.g., Tibshirani, 1996), five out eight regressors appear to be not related to the response; however, unlike the L_1 penalty, the qL_1 framework provides confidence intervals which typically are quite informative.

TABLE 1. Parameter estimates, standard errors and Normal-based 95% confidence intervals via the qL_1 penalized regression ($c = 0.1$).

			95% CI	
covariate	Est	SE	inf	sup
lcavol	0.577	0.0927	0.395	0.759
lweight	0.227	0.0797	0.071	0.383
age	−0.063	0.0480	−0.157	0.031
lbph	0.083	0.0486	−0.013	0.178
svi	0.229	0.0894	0.054	0.404
lcp	0.003	0.0476	−0.090	0.097
gleason	0.036	0.0456	−0.054	0.125
pgg45	0.059	0.0448	−0.029	0.147

To illustrate the performance of the qL_1 we present results from a small simulation experiment ($n = 100$ with 2 predictors out $p = 20$ covariates). We consider the standard LASSO with tuning parameter selected by 10-fold CV and the qL_1 with λ selected by GCV using the trace of the hat matrix to quantify the model dimension. Figure 3 shows two sampling distributions for the estimator of a null parameter; unlike L_1, qL_1 appears to ensure Normal distribution for $\hat{\beta}$.

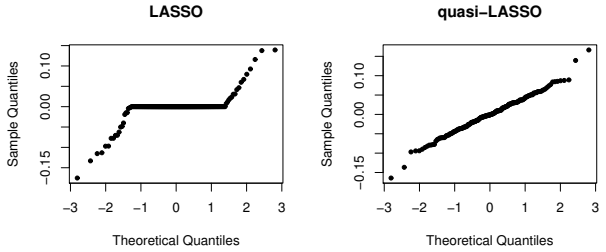


FIGURE 3. Normal Q-Q plots of sampling distributions of LASSO and quasi-LASSO estimator $\hat{\beta}$ when $\beta = 0$.

Table 2 shows the mean, standard deviation and average of the estimated SE for 4 parameter estimators; sandwich formula in the qL_1 framework appears to work reasonably well.

TABLE 2. Mean and standard deviation of the sampling distributions for 4 parameter estimators in the simulation study; $\bar{\text{SE}}$ indicates the averages of the standard errors obtained in each replicate (in each block the 4 columns refer respectively $\beta = 0.5, -0.5, 0, 0$).

	LASSO				quasi-LASSO			
m($\hat{\beta}$)	0.374	−0.358	0.000	0.001	0.390	-0.374	-0.001	0.003
sd($\hat{\beta}$)	0.116	0.100	0.036	0.036	0.109	0.096	0.051	0.050
$\bar{\text{SE}}$	0.103	0.103	0.107	0.108	0.099	0.098	0.048	0.047

4 Discussion

We have presented a smooth parametric approximation for the LASSO penalty. The resulting quasi-LASSO regression may be implemented via iterative least squares and guarantees normality of the parameter estimator, even in presence of zero coefficients. A sandwich formula is available and allows to build reliable normal-based confidence intervals. Smooth approximations of the L_1-norm have been discussed by Osborne *et al.* (2000); however their approximations, different from the (1), are employed to motivate some formulas for the covariance matrix and not for estimation. The proposed qL_1 approximation depends upon a threshold value $c > 0$. As c increases qL_1 reduces to L_2 (i.e. ridge regression), and 'sparsity' is lost. Therefore selection of c is an important issue of the proposed approach which deserves major investigation. Another important issue refers to the computation of the model *df* which may be computed by the number of 'non-zero' estimates, namely $\#\{|\hat{\beta}_j| > c\}$, or alternatively it is possible use the trace of the 'hat matrix', $df = \mathrm{tr}(X^T X H^{-1})$. In conclusion, we believe the proposed approach represents a possible alternative to get reliable inference (confidence intervals and p-value) within the frequentist framework as an alternative to the Bayesian one (Kyung et al., 2010).

References

Knight K. and Fu W. (2000) Asymptotics for lasso-type estimators. *Annals of Statistics*, **28**, 1356-1378.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004) Least angle regression. *Annals of Statistics*, **32**, 407–489.

Kyung, M., Gilly, J., Ghoshz, M.,and Casella, G. (2010) Penalized Regression, Standard Errors, and Bayesian Lassos. *Bayesian Analysis*, **5**, 1–44.

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). On the LASSO and its dual. *Journal of Computational Graphical Statistics*, **9**, 319–337.

Pötscher, B. M., and H. Leeb (2009) On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *Journal of Multivariate Analysis*, **10**, 2065–2082.

Tibshirani (1996) Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society B*, **58**, 267–288.

Zou, H., Hastie, T., and Tibshirani, R. (2007) On the 'degrees of freedom' of the lasso. *Annals of Statistics*, **35**, 2173–2192.

Modelling random effects using GAMLSS

Graciela-Muniz Terrera[1], Ardo van den Hout[1], Mikis Stasinopoulos[2], Bob Rigby[2]

[1] MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 OSR, UK. E-mail: graciela.muniz@mrc-bsu.cam.ac.uk
[2] London Metropolitan University, UK

Abstract: An important assumption in many random-effects models is that the response variable is normally distributed. This assumption is not always appropriate. To describe cognitive decline in the older population, a semi-parametric generalised additive model for location, scale and shape (GAMLSS) was fitted in which the response variable was assumed to have a non-normal distribution. Data from the Cambridge city over 75 Cohort Study, a longitudinal study of ageing, is used to illustrate how model fit was improved by relaxing the normality assumption.

Keywords: cognition; longitudinal data; beta-binomial distribution.

1 Introduction

Many regression-based mixed-effects models assume that the response variable is normally distributed. In a longitudinal study of ageing, it is of interest how the response variable measuring cognitive decline can be regressed on changing age and other covariates. To investigate this, a mixed-effects model can be applied where the random effects structure takes into account that longitudinal measurements within individuals are correlated. Given a fitted model, inspecting the distribution of the residuals raises the question whether the fit can be improved by replacing the normal distribution of the response variable by another distribution.

Generalised additive models for location, scale and shape (GAMLSS, Rigby and Stasinopoulos, 2005) are semi-parametric regression-based models designed to overcome some of the limitations of generalised linear models (GLM, Nelder and Wedderburn, 1972), generalised additive models (GAMS, Hastie and Tibshirani, 1990) and linear mixed models (LME, Laid and Ware, 1982). GAMLSS relaxes the exponential family distribution assumption in GLMs and GAMs, and it allows the modelling of skewness and kurtosis in a way that goes beyond the standard LME framework.

The (random-effects) GAMLSS model is defined as follows. Let y_i, $i = 1, ..., n$, be conditional independent observations with density function $f(y_i|\theta^i)$, where $\theta^i = (\mu_i, \sigma_i, \nu_i, \tau_i)$ is a vector of four distribution parameters. The

four parameter vectors, each with dimension $n \times 1$, can be linked to explanatory variables by

$$\begin{aligned} g_1(\mu) &= \eta_1 = \mathbf{X}_1\beta_1 + \sum_{j=1}^{J_1} \mathbf{Z}_{j1}\gamma_{j1} \\ g_2(\sigma) &= \eta_2 = \mathbf{X}_2\beta_2 + \sum_{j=1}^{J_2} \mathbf{Z}_{j2}\gamma_{j2} \\ g_3(\nu) &= \eta_3 = \mathbf{X}_3\beta_3 + \sum_{j=1}^{J_3} \mathbf{Z}_{j3}\gamma_{j3} \\ g_4(\tau) &= \eta_4 = \mathbf{X}_4\beta_4 + \sum_{j=1}^{J_4} \mathbf{Z}_{j1}\gamma_{j4}, \end{aligned}$$

where $g_k(\cdot)$ are known link functions, $\mathbf{X}_k$ and $\mathbf{Z}_{jk}$ are known design matrices. In the normal random effects GAMLSS model the γ_{jk} are random variables assumed to be normally distributed with zero mean vector and variance-covariance matrix $G_{jk}^{-1}(\lambda_{jk})$ depending on hyper-parameters λ_{jk}, for $j = 1, 2, \ldots, J_k$, where $k = 1, 2, 3, 4$. The formulation with $\mathbf{Z}_{jk}$ and γ_{jk} allows combinations of different types of additive random-effects terms. For fixed hyper parameters λ_{jk} the GAMLSS algorithm finds the posterior mode (MAP) estimators for the fixed effects β-parameters and the random effects γ-parameters respectively, see Rigby and Stasinopoulos (2005). Different methods of estimating the hyper-parameters were discussed briefly in Rigby and Stasinopoulos (2005) but their computational efficiency was never investigated. The methods involve using i) an EM algorithm ii) a Laplace approximation of the marginal likelihood of the hyper-parameters and iii) a local maximum likelihood performed at each iteration of the GAMLSS algorithm. The final aim of the current work is to compare and evaluate these methods.

However, as an intermediate step, this paper will investigate GAMLSS models where the random effects distribution is modelled as a discrete distribution estimated by non-parametric maximum likelihood.

The investigation is motivated by data from the Cambridge City over 75 Cohort Study, a UK population-based longitudinal study of ageing. Initially, a representative sample of 2086 individuals (65% women) aged at least 75-years-old living in the Cambridge area who were registered at seven primary care practices in the Cambridge City area were invited to participate in the study. After baseline, further interviews were conducted on average 2, 7, 9, 12, 17 and 21 years later. Cognitive function was measured using the Mini Mental State Examination (MMSE) with integer scores between 0 and 30 with high values indicating good cognition. Its distribution is known to be skewed to the left.

2 Models

Where a random-effect is included in the linear predictor of a GLM, an assumed standard normal distribution for the random-effect is approximated by Gaussian quadrature with K mass-points and the resulting likelihood is a likelihood of a finite mixture with known mixture proportions p_k at known mass-points z_k, $k = 1, .., K$ (Aitkin, 1996a).
A possible next step is to consider the mixture proportions p_k and mass-points z_k to be unknown parameters. Aitkin (1996a) discusses non-parametric maximum likelihood estimation (NPML) of these parameters along with the fixed-effects parameters of the GLM.
We will discuss two GAMLSS models for cognitive decline with non-parametric random-effects. Model A assumes the normal distribution for the response variable and the location and scale parameters of this distribution are regressed as follows

$$\begin{aligned} g_1(\mu_{it}) &= \mu_{it} = \beta_{1i} + \beta_{2i} Age_{it} + \beta_3 Age_{it}^2 + \beta_4 Sex_i \\ g_2(\sigma) &= \log(\sigma) = \beta_\sigma. \end{aligned}$$

where β_{1i} and β_{2i} are the random effects, and where Age_{it} is individual i's age minus 75 year at interview t, for $t = 1, ..., n_i$, and Sex is a binary variable (with values 1 for women, 0 for men).
Using non-parametric random effects means effectively using a discrete bivariate distribution for $(\beta_{1i}, \beta_{2i})^\top$. This distribution is determined by mixture proportions π_k, $k = 1, 2, \ldots, K$, and bivariate mass-points given by $u_k = (u_{1k}, u_{2k})$, $k = 1, 2, \ldots, K$. The u_k's and π_k's are estimated using maximum likelihood estimation.
Model B assumes the beta-binomial distribution for the response variable. Given that MMSE scores are integers from 0 up to 30, a discrete distribution for the response variable is more appropriate. In Model B, the location and scale parameters of this distribution are regressed as follows

$$\begin{aligned} g_1(\mu_{it}) &= \text{logit}(\mu_{it}) = \alpha_{1i} + \alpha_{2i} Age_{it} + \alpha_3 Age_{it}^2 + \alpha_4 Sex_i \\ g_2(\sigma) &= \log(\sigma) = \alpha_\sigma. \end{aligned}$$

In the above, we follow Muniz-Terrera *et al.* (2008) with respect to including the quadratic term for age. The beta-binomial is not in the natural exponential family. Even if the random effects are removed, model B is not a GLM.
The models above can readily be extended within the GAMLSS framework. For example, in both models we could include covariates in the regression for the scale parameters.

3 Estimation

In the normal random effects GAMLSS model, for fixed random effects hyper parameters, the β and γ parameters in GAMLSS models are fitted

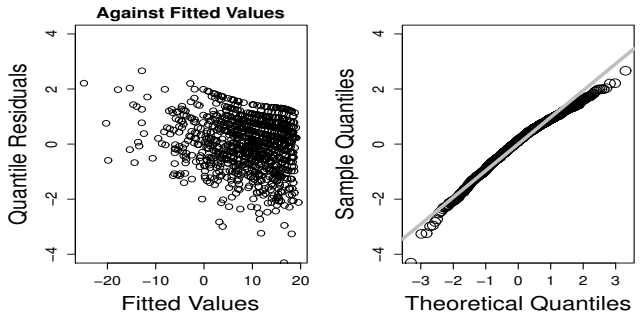


FIGURE 1. Normalised quantile residuals for GAMLSS Model A with the normal distribution for the response variable.

by maximising the penalised likelihood function

$$l_p = \sum_{i=1}^{n} \log\left(f(y_i|\theta^i)\right) - \frac{1}{2}\sum_{jk} \gamma_{jk}^{\top}\mathbf{G}_{jk}\gamma_{jk},$$

where symmetric matrices $\mathbf{G}_{jk}$ parameterise the independent (prior) normal distributions of the random effects γ_{jk} (Rigby and Stasinopoulos, 2005). We will skip details here, but it is worth mentioning the important role of a backfitting algorithm in the estimation of the additive predictor $\mathbf{Z}_{jk}\gamma_{jk}$. Using backfitting makes it relatively easy to add additive terms to the linear predictors.

Models A and B are estimated by non-parametric maximum likelihood using the EM algorithm as described in Aitkin (1996a). Within this algorithm, the standard GAMLSS methods are used. Hence, it is possible to investigate different distributions for the response (e.g., normal, t, binomial, or beta-binomial). The reason for considering both continuous and discrete response distributions is that, although the response variable is strictly a discrete variable, the counts are moderately large. The models can be estimated in R with the function gamllsNP() that can be found in the package gamlss.mx.

4 Results

There are 2086 individuals in the data sets with a total of 4450 observations. The maximum number of observations per individual is 7. There are 924 individuals with only one observation. In the following we work with a random subset of 500 individuals. With respect to the number of observations 1 up to 7, the frequencies in this subset are 225, 128, 66, 50, 20, 9, and 2.

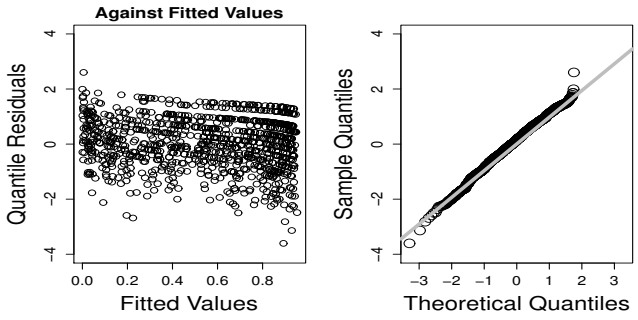


FIGURE 2. Normalised quantile residuals for GAMLSS Model B with the beta-binomial distribution for the response variable.

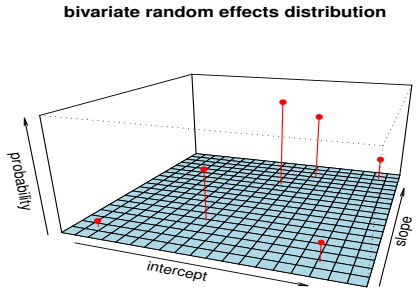


FIGURE 3. The bivariate random effects distribution.

Figure 1 shows the normalised quantile residuals for Model A (with $K = 6$), and the Q-Q plot. The residuals should have a standard normal distribution if the model is correct. The fit is not bad, but there is an undesired effect due to the upper bound for the scores of the response variable MMSE. The upper bound is not taken into account when a normal distribution is assumed for the MMSE. The number of mass points $K = 6$ in Model A was chosen by comparing AICs for different choices of K.

Figure 2 shows the normalised quantile residuals for Model B with the beta-binomial distribution and $K = 6$. Note that there is no systematic effect for the residuals for fitted values close to 30. Comparing AICs, Model B also shows a better fit than Model A (5390.3 vs. 5647.8).

Figure 3 shows the bivariate random effects distribution. Figure 4 shows the empirical Bayes predictions (Aitkin, 1996b) and the 6 fitted component curves of the means of the $K = 6$ classes for model B.

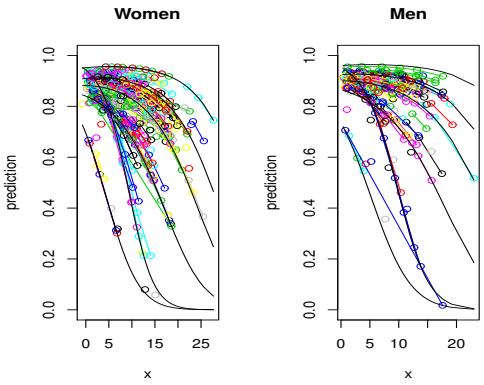


FIGURE 4. Empirical Bayesian predictors for the mean for Model B with the beta-binomial distribution for the response variable plotted against age. The means of the six classes are given by the black curves.

References

Aitkin, M. (1996a). A general maximum likelihood analysis of overdispersion in generalized linear models, *Statistics and Computing* **6**, 251-262.

Aitkin, M. (1996b). Empirical Bayes shrinkage using posterior random effect means from nonparametric maximum likelihood estimation in general random effect models. *Statistical Modelling: Proceedings of the 11th IWSM 1996*, 87-94.

Hastie, T.J., and Tibshirani, R.J. (1990). *Generalized additive models*, London: Chapman and Hall.

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

Nelder J. Wedderman R. (1972).Generalized Linear Models. *Journal of the Royal Statistical Society. Series A* **135**, 370384

Muniz-Terrera, G., Matthews, F.E., and Brayne, C. (2008). A comparison of parametric models for the investigation of the shape of cognitive change in the older population, *BMC Neurology* **8**.

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

Least Orthogonal Distance Estimator for SEM based on Singular Value Decomposition

A. Naccarato[1], D. Zurlo[2]

[1] Roma Tre University faculty of Economics, via Silvio D'Amico 77, Rome, naccarat@uniroma3.it
[2] Roma Tre University faculty of Economics, via Silvio D'Amico 77, Rome, dzurlo@uniroma3.it

Keywords: SEM; Principal Component; Singular Value Decomposition;.

1 Introduction

The aim of this paper is to present a consistent full information estimator of structural parameters in a SEM, based on the singular value and vector of a matrix deriving from the so-called over-identifying restriction. The starting idea whose development has given rise to Least Orthogonal Distance Estimator(LODE) can be tracked back to the work of (Pieraccini, 1969), in which 2SLS were obtained as generalized least square estimator applied to the system of so called identifying restriction; the result was afterwards extended to 3SLS (Pieraccini,1978). With this in mind and making reference to the work of Pearson's "Lines and planes of closest fit" (Pearson 1901) the LODE method of estimation has been derived under the consideration that the identifying system is nothing else but linear relations between variables affected by error (Naccarato and Pieraccini, 2008).
The original form of LODE was based on characteristic roots and vectors. Subsequent developments led to the use of the Singular Value Decomposition instead of the Spectral Decomposition. This is because an algorithm based on SVD is numerically more robust compared to an algorithm based on SD, in the sense of algorithm implementation, where robustness refers to the algorithm's probability of converging (Markovsky and Van Huffel, 2007). This leads to greater stability of LODE estimates. Another peculiarity of the LODE estimator is that this method doesn't impose a prior choice of the dependent variable in the system's equation.

2 The estimator

Making use of standard notations, the structural form of a simultaneous equations model can be defined as follows :

$$\underset{n,m}{Y}\ \underset{m,m}{\Gamma} + \underset{n,k}{X}\ \underset{k,m}{B} + \underset{n,m}{U} = \underset{n,m}{0} \tag{1}$$

where Y is the matrix $n \times m$ of endogenous variables, Γ is the corresponding $m \times m$ matrix of structural parameters, X is the matrix $n \times k$ of exogenous variables and B is the matrix $k \times m$ of their structural parameters. Finally U is the $n\times$ m matrix of disturbances for which standard hypotheses are supposed to hold:

$$\begin{array}{c} E(vecU) = 0 \\ E(vecU(vecU))^T = \Omega \otimes I \end{array} \tag{2}$$

Under non singularity condition for Γ the reduced form of the equations is derived as:

$$\underset{n,m}{Y} = \underset{n,kk,m}{X}\,\underset{}{\Pi} + \underset{n,m}{V} \tag{3}$$

where:

$$\begin{array}{l} \underset{k,m}{\Pi} = -\underset{k,mm,m}{B}\,\Gamma^{-1} \\ \underset{k,m}{V} = -\underset{k,mm,m}{U}\,\Gamma^{-1} \end{array} \tag{4}$$

Post-multiplying by Γ the first equation in (4) we obtain

$$\underset{k,mm,m}{\Pi}\ \underset{}{\Gamma} = -\underset{k,m}{B} \tag{5}$$

which represents the relation between reduced and structural form. Since this is a system of k equation with $m \times (m + k)$ unknowns, exclusion constraints are introduced in each equation in order to find the solution with respect to Γ and B. In terms of Π, the result is this system

$$\begin{cases} \underset{k_{1i},1}{\hat{\pi}^i_{01}} = \underset{k_{1i},m_{1i}-1}{\hat{\Pi}^i_{11e}}\ \underset{m_{1i}-1,1}{\Gamma_{1i}} + \underset{k_{1i},1}{B_{1i}} + \underset{k_{1i},1}{\varepsilon_{1i}} \\ \underset{k_{2i},1}{\hat{\pi}^i_{02}} = \underset{k_{2i},m_{1i}-1}{\hat{\Pi}^i_{12e}}\ \underset{m_{1i}-1,1}{\Gamma_{1i}} + \underset{k_{2i},1}{\varepsilon_{2i}} \end{cases} \tag{6}$$

This is the identifying system of the i-th equation, where $\hat{\pi}^i_{01}$ and $\hat{\pi}^i_{02}$ are the OLS estimates of coefficients related to endogenous dependent variable and to exogenous variables respectively included and excluded, while $\hat{\Pi}^i_{11e}$ and $\hat{\Pi}^i_{12e}$ refer to the other endogenous variables included and to exogenous variables respectively included and excluded. Γ_{1i} are the coefficients of m_{1i} endogenous variables included in each system's equation and B_{1i} are the k_{1i} parameters related to the exogenous variable included.
For the whole system the second equation of (6) will be

$$\underset{r,m}{\hat{\pi}_2} = \underset{r,z-mz-m,m}{\hat{\Pi}_2}\ \underset{}{\Gamma_1} + \underset{r,m}{E_2} \quad z = \sum_{i=1}^{m} m_{1i} \;\; r = \sum_{i=1}^{m} k_{2i} \tag{7}$$

with

$$\hat{\pi}_2 = \begin{bmatrix} \hat{\pi}_{21} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \hat{\pi}_{2m} \end{bmatrix} \quad \hat{\Pi}_2 = \begin{bmatrix} \hat{\Pi}_{21} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \hat{\Pi}_{2m} \end{bmatrix} \quad \Gamma_1 = \begin{bmatrix} \Gamma_{11} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Gamma_{1m} \end{bmatrix}$$

and

$$E(E_2 E_2^T) = \underset{m,m}{\Omega} \otimes \left(\underset{k_2,k_2}{R_{22}} \right)^{-1}$$

where R_{22} comes from

$$\underset{k,k}{(X^T X)^{-1}} = \begin{bmatrix} \underset{k_1,k}{R_1} \\ \underset{k_2,k}{R_2} \end{bmatrix} = \begin{bmatrix} \underset{k_1,k_1}{R_{11}} & \underset{k_1,k_2}{R_{12}} \\ \underset{k_1,k_1}{R_{12}} & \underset{k_2,k_2}{R_{22}} \end{bmatrix}$$

Equation (7) can be viewed as a linear equation between variables subject to error. These variables are the OLS estimates of π_2 and Π_2, so to estimate Γ_1 the idea is to use an estimator based on least orthogonal distance. Equation (7) can be written as

$$\underset{r,r}{Q} \begin{bmatrix} \underset{rm}{\hat{\pi}_2}; & \underset{r,z-m}{\hat{\Pi}_2} \end{bmatrix} \begin{bmatrix} \underset{}{-I_m} \\ \underset{z-m,m}{\Gamma_1} \end{bmatrix} = \underset{r,rr,m}{Q\, E_2} \text{ with } Q^T Q = \underset{m,m}{\hat{\Omega}}^{-1} \otimes \left(\underset{k_2,k_2}{R_{22}} \right)$$

where $\hat{\Omega}$ comes from limited information estimates. The rank of the matrix $T = \underset{r,r}{Q} \begin{bmatrix} \underset{rm}{\hat{\pi}_2}; & \underset{r,z-m}{\hat{\Pi}_2} \end{bmatrix}$ is z. To be able to write π_2 as a function of Π_2 we have to reduce this matrix to a matrix of rank $z - m$. From the Eckart-Young matrix approximation theorem we know that the matrix of rank $z-m$ that has the least orthogonal distance from T is

$$\underset{r,m}{A(T)} = \underset{r,rr,mm,m}{U\ \Lambda\ V^T}$$

U and V are the matrices of left and right singular vectors of T, Λ is equal to the diagonal matrix of singular value of T except for the last m element equal to zero. Given that

$$A(T) \begin{bmatrix} \underset{m,m}{V_{12}} \\ \underset{z-m,m}{V_{22}} \end{bmatrix} = 0$$

where $\begin{bmatrix} \underset{m,m}{V_{12}} \\ \underset{z-m,m}{V_{22}} \end{bmatrix}$ are the last $z-m$ right singular vectors of T which belong to the null space of $A(T)$, then the least orthogonal distance estimates of Γ_1 will be

$$\begin{bmatrix} \underset{}{-I_m} \\ \underset{z-m,m}{\hat{\Gamma}_1} \end{bmatrix} = \begin{bmatrix} \underset{m,m}{V_{12}} \\ \underset{z-m,m}{V_{22}} \end{bmatrix} \begin{bmatrix} \underset{mm}{-V_{12}^{-1}} \end{bmatrix}$$

The estimates of B_1 can easily come from applying OLS on this equation

$$Y_1 \hat{\Gamma}_1 = X_1 B_1 + U \tag{8}$$

and so

$$\hat{B}_1 = (X_1^T X_1)^{-1} X_1^T (Y_1 \hat{\Gamma}_1) \tag{9}$$

The consistency of FI LODE estimates is proven in Zurlo(2010) and no assumption is made on the structural form error component distribution. Futhermore a very extensive simulation experiment (Naccarato and Zurlo, 2007) and Zurlo (2010) shows the good performances of this method in terms of unbiasedness and MSE, comparing with classical SEM estimator, especially for small sample size of the simulating data.

References

Golub G. H. and Van Loan C. (1980)*An analysis of the total least square problem*, SIAM J. Number Anal. 17

Markovsky I.and Van Huffel S. (2007).*Overview on total least square* Signal Processing vol 87 pp 2283-2303 Oxford: Clarendon Press.

Naccarato A. and Zurlo D. (2008), *A Monte Carlo Study on Full Information Methods in Simultaneous Equation Models*, Quaderni di Statistica vol. 10

Naccarato A. (2007) *Full Information Least Orthogonal Distance*, Quaderni di Statistica vol. 10

Naccarato and Pieraccini (2008) *a Full information least distance estimator* Statistica vol. spec. pp52-76

Zurlo (2010) *the class of Least Orthogonal Distance Estimator of structural parameters for SEM* Phd thesis Dept. of Economics, Roma Tre University

A comparison of methods for factor analysis of multivariate geostatistical data

Samuel D. Oman[1], Bella Vakulenko-Lagun[1], Michael Zilberbrand [2]

[1] Department of Statistics, Hebrew University, Mount Scopus, Jerusalem 91905 Israel

[2] Research Division, Hydrological Service of Israel, P.O.B. 36118, Jerusalem 91360 Israel

Abstract: We describe four methods for factor analysis of spatially correlated vectors: the Linear Model of Coregionalization and three recently proposed alternatives. We apply the three methods to analyze the concentrations of nine chemicals in water samples taken from 306 springs. The methods give quite different results, with those of the Linear Model of Coregionalization being much more interpretable. We suggest some possible explanations for this, which may be relevant in other applications as well

Keywords: Factor analysis; Principal component analysis; Semivariogram matrix; Spatial correlation.

1 Introduction

Many geostatistical applications involve spatially dependent vectors, observed at different sites in a sampling region. Examples are concentrations of different chemicals in topsoil samples (Webster et al, 1994), and percent cover of different species on line transects or in subplots (Maestre et al, 2005). It is often useful to perform a principal component or factor-analytic type of analysis, in order to understand and model both the correlations within the vectors and the structure of their spatial correlation. In this paper we briefly describe four such methods: the Linear Model of Coregionalization (LMC; Section 2), and three recently proposed alternatives (Section 3). In Section 4, we apply them to data on chemical concentrations in water samples taken from 306 springs. The methods give quite different results, with those of the Linear Model of Coregionalization being much more interpretable. In Section 5 we offer a possible explanation for the difference in interpretability.

If $\mathbf{Y}(\mathbf{s})$ denotes the vector of observations obtained at site $\mathbf{s}$ in a sampling region D, all of the methods assume that $\{\mathbf{Y}(\mathbf{s}) : \mathbf{s} \in D\}$ is a second-order stationary m-variate process. We define $\mathbf{\Sigma} = \mathrm{cov}(\mathbf{Y}(\mathbf{s}))$, let $\mathbf{C}(\mathbf{h}) = \mathrm{cov}(\mathbf{Y}(\mathbf{s}), \mathbf{Y}(\mathbf{s}+\mathbf{h}))$, and let $\mathbf{\Gamma}(\mathbf{h}) = (1/2)E[\mathbf{Y}(\mathbf{s}+\mathbf{h}) - \mathbf{Y}(\mathbf{s})][\mathbf{Y}(\mathbf{s}+\mathbf{h}) -$

$\mathbf{Y}(\mathbf{s})]^t$ denote the semivariogram matrix of direct and cross semivariograms for all components Y_i and Y_j of $\mathbf{Y}$ observed at a lag $\mathbf{h}$ apart.

2 The Linear Model of Coregionalization

The LMC (Journel and Huijbregts, 1978; Wackernagel, 1989; Goovaerts, 1994) assumes that

$$\mathbf{\Gamma}(\mathbf{h}) = \gamma_1(\mathbf{h})\mathbf{V}_1 + \cdots + \gamma_p(\mathbf{h})\mathbf{V}_p, \tag{1}$$

where each $\mathbf{V}_j$ is a non-negative definite matrix of sills. The γ_j are basic univariate semivariogram functions (typically of an assumed parametric form) which describe covariation over different ranges θ_j; their number (often, two or three) and form are typically based on subject-matter considerations and a preliminary analysis of the data.
To estimate the $\mathbf{V}_j$ and θ_j, we first compute empirical semivariogram matrices $\hat{\mathbf{\Gamma}}(\mathbf{h}_1), \ldots, \hat{\mathbf{\Gamma}}(\mathbf{h}_n)$ at lags $\mathbf{h}_1, \ldots, \mathbf{h}_n$, and then minimize

$$\sum_{i=1}^{n} w_i \|\gamma_1(\mathbf{h}_i)\mathbf{V}_1 + \cdots + \gamma_p(\mathbf{h}_i)\mathbf{V}_p - \hat{\mathbf{\Gamma}}(\mathbf{h}_i)\|_F^2. \tag{2}$$

Here, $\|A\|_F^2 = \mathrm{tr}(\mathbf{A}\mathbf{A}^t)$ denotes the Frobenius norm of a square matrix $\mathbf{A}$ and the w_i are appropriately chosen weights. The minimization with respect to the sill matrices, which requires that they be non-negative definite, is done using the algorithm of Goulard and Voltz (1992), which is known to converge to a unique solution (Oman and Vakulenko-Lagun, 2009).
If each $\mathbf{V}_j = \mathbf{A}_j\mathbf{A}_j^t$, (1) is equivalent to assuming

$$\mathbf{Y}(\mathbf{s}) = \sum_{j=1}^{p}\sum_{k=1}^{m} \mathbf{a}_{jk} z_{jk}(\mathbf{s}), \tag{3}$$

where $\{\mathbf{a}_{jk}\}$ are the columns of $\mathbf{A}_j$, the $\{z_{jk}\}$ are univariate spatial processes which are uncorrelated with one another, both pointwise and spatially, and the semivariogram function of z_{jk} is γ_j. Thus, for each j the z_{jk} may be interpreted as independent factors acting at a spatial scale characterized by γ_j, with $\{\mathbf{a}_{jk}\}$ giving the loadings of $\mathbf{Y}$ on z_{jk}. Often, as in factor analysis, the first few $\mathbf{a}_{jk}$ (ordered according to length) are rotated, to make the z_{jk} more interpretable. We remark that the factors z_{jk} are not given explicitly in terms of $\mathbf{Y}$ (unless $p = 1$, the intrinsic LMC, in which case (3) can be inverted). Thus, factor scores must be obtained by cokriging (Wackernagel, 1994).

3 Some Alternative Methods

In part motivated by the above-mentioned problems with factor scores, Bailey and Krzanovski (2000) and Krzanovski and Bailey (2007) proposed

three alternative approaches, designed to obtain factors as explicit linear combinations of the Y_j. The first seeks to obtain factors with zero cross-correlations at given lags; the second seeks factors with given autocorrelation functions; and the third seeks both objectives.

3.1 Identifying Factors with Zero Cross-Correlations

To obtain m factors with negligible cross-correlations at given lags $\mathbf{h}_1, \ldots, \mathbf{h}_n$, if $\hat{\mathbf{\Sigma}} = \mathbf{F}\mathbf{F}^t$ let $\mathbf{Q}_i = \mathbf{F}^{-1}\hat{\mathbf{\Gamma}}(\mathbf{h}_i)\mathbf{F}^{-t}$. The proposal is then to find an orthogonal matrix $\mathbf{U}$ which minimizes

$$\sum_{i=1}^{n} \|\mathbf{U}^t\mathbf{Q}_i\mathbf{U} - \mathrm{diag}(\mathbf{U}^t\mathbf{Q}_i\mathbf{U})\|_F^2, \tag{4}$$

and define the vector of m factors by $\mathbf{Z}(\mathbf{s}) = \mathbf{U}^t\mathbf{F}^{-1}\mathbf{Y}(\mathbf{s})$. Since the estimated covariance and variogram matrices of $\mathbf{Z}$ are $\hat{\mathbf{\Sigma}}_{\mathbf{Z}} = \mathbf{I}$ and $\hat{\mathbf{\Gamma}}_{\mathbf{Z}}(\mathbf{h}_i) = \mathbf{U}^t\mathbf{Q}_i\mathbf{U}$, we see that if the minimum of (4) is zero, the components of $\mathbf{Z}$ have zero cross-correlations at the given lags. Minimization may be done using the Common Principal Component Analysis (CPC) algorithm of Clarkson (1988a, b); since zero is usually not achieved, the cross-correlations are only approximately zero. Nothing can be said about the autocorrelation functions of the factors.

3.2 Identifying Factors Acting at Different Scales

To find factors $z_k(\mathbf{s}) = \mathbf{a}_k^t\mathbf{Y}_(\mathbf{s})$ with fitted parametric autocorrelation functions $\rho(\cdot|\hat{\theta}_k)$, the method is as follows. Select a set of lags $\mathbf{h}_1, \ldots, \mathbf{h}_n$, and then sequentially minimize

$$\frac{\sum_i[(\mathbf{a}^t\mathbf{\Gamma}^*(\mathbf{h}_i)\mathbf{a})/(\mathbf{a}^t\mathbf{\Sigma}^*\mathbf{a}) - (1 - \rho(\mathbf{h}_i|\theta))]^2}{\sum_i[1 - (\mathbf{a}^t\mathbf{\Gamma}^*(\mathbf{h}_i)\mathbf{a})/(\mathbf{a}^t\mathbf{\Sigma}^*\mathbf{a})]^2} \tag{5}$$

with respect to $\mathbf{a}$ and θ. In the first minimization, $\mathbf{\Gamma}^*$ and $\mathbf{\Sigma}^*$ are estimates using the original $\mathbf{Y}(\mathbf{s}_i)$; in the subsequent minimizations, they are computed from projections of the $\mathbf{Y}(\mathbf{s}_i)$ onto the subspace orthogonal to the span of the $\mathbf{a}_j$ already computed. The denominator in (5) is to help obtain a maximal range θ, by preventing $(\mathbf{a}^t\mathbf{\Gamma}^*(\mathbf{h}_i)\mathbf{a})/(\mathbf{a}^t\mathbf{\Sigma}^*\mathbf{a})$ from being too close to 1 (which would lead to a small value of θ in the numerator). This hopefully will ensure that the factors can be ranked by their ranges $\theta_1 \geq \cdots \geq \theta_K$. Nothing can be said about the cross-correlations between the factors. We remark that although the $\mathbf{a}_j$ are orthogonal, the z_j are not orthogonal factors in that $\mathrm{cov}(z_i, z_j)$ need not equal zero for $i \neq j$.

3.3 A Synthesis

To obtain the objectives of Sections 3.2 and 3.3, the proposal is an alternating iterative procedure. At iteration $j+1$, the two steps are:
(a) Using the CPC algorithm, find an orthogonal matrix $\mathbf{U} = \mathbf{U}_{j+1}$ which minimizes

$$\sum_{i=1}^{n} \|\mathbf{U}^t(\mathbf{Q}_i - \mathbf{U}_j \boldsymbol{\Delta}_i^{(j)} \mathbf{U}_j^t)\mathbf{U} - \mathrm{diag}(\mathbf{U}^t[\mathbf{Q}_i - \mathbf{U}_j \boldsymbol{\Delta}_i^{(j)} \mathbf{U}_j^t]\mathbf{U})\|_F^2. \quad (6)$$

(b) For each component $k = 1, \ldots, m$, compute $\hat{\theta}_k^{(j+1)}$ by fitting $\gamma(\cdot|\theta_k)$ to the k-th diagonal elements of $\mathbf{U}_{j+1}^t \mathbf{Q}_1 \mathbf{U}_{j+1}, \ldots, \mathbf{U}_{j+1}^t \mathbf{Q}_n \mathbf{U}_{j+1}$, and then let $\boldsymbol{\Delta}_i^{(j+1)} = \mathrm{diag}(\gamma(\mathbf{h}_i|\hat{\theta}_1^{(j+1)}), \ldots, \gamma(\mathbf{h}_i|\hat{\theta}_m^{(j+1)}))$.
Note that if there is convergence, $\mathbf{U}_{j+1} \approx \mathbf{U}_j$ and thus (6) is close to $\sum_{i=1}^n \|\mathbf{U}_{j+1}^t \mathbf{Q}_i \mathbf{U}_{j+1} - \mathrm{diag}(\mathbf{U}_{j+1}^t \mathbf{Q}_i \mathbf{U}_{j+1})\|_F^2$; *i. e.*, we have approximately obtained zero cross-correlations at the specified lags.
If the off-diagonal elements of $\mathbf{U}_j^t \mathbf{Q}_i \mathbf{U}_j$ are negligible, then from (b) $\boldsymbol{\Delta}_i^{(j)} \approx \mathrm{diag}(\mathbf{U}_j^t \mathbf{Q}_i \mathbf{U}_j) \approx \mathbf{U}_j^t \mathbf{Q}_i \mathbf{U}_j$. Thus $\mathbf{U}_j \boldsymbol{\Delta}_i^{(j)} \mathbf{U}_j^t \approx \mathbf{Q}_i$, so subtracting it in (6) should cause the objective function to decrease with j, helping to ensure convergence.

4 Application

We now analyze the chemical composition of water samples taken from 306 springs in a 25×70 km mountainous region of Israel. For each sample, the concentrations (mg/l) of NO_3, HCO_3, SO_4, Cl, K, Na, Mg and Ca were measured. As the distributions are quite skewed, we work in log units.
A principal components analysis of the sample covariance matrix gives three components which explain 95% of the variability. From their loadings, they can be interpreted as anthropogenic factors such as cowsheds and domestic sewage (high levels of NO_3 and K); a contrast between seawater salinity (Cl, Na, Mg, SO_4) and NO_3; and (high levels of HCO_3, Ca) carbonate (aquifer rock) dissolution.
The semivariograms of all the components show large nugget effects, with a maximum range of approximately 20 km; fitting a LMC with a nugget and spherical autocorrelation function gives an estimated range of 19.58 km. 95% of the sum of the eigenvalues of the two sill matrices is explained by seven components, whose loading vectors (following a varimax rotation with Kaiser normalization) are shown in Table 1.
The nugget matrix components are essentially the same as for the covariance matrix: anthropogenic factors, a combination of anthropogenic factors and seawater salinity, and a combination of anthropogenic factors and carbonate dissolution. For the spherical sill matrix, the first component is NO_3, the second is essentially a combination of anthropogenic factors

TABLE 1. Rotated eigenvectors explaining 95% of the variation of the LMC sill matrices, with their squared lengths (λ).

	Nugget				Spherical		
λ	0.270	0.192	0.180	0.058	0.157	0.131	0.026
NO_3	**-0.170**	**0.411**	**-0.237**	**-0.107**	**0.380**	**-0.153**	0.030
HCO_3	-0.008	0.005	-0.018	**-0.089**	0.002	-0.004	0.065
SO_4	-0.034	0.027	**-0.181**	-0.030	0.011	**-0.109**	0.076
Cl	-0.049	0.042	**-0.149**	-0.044	0.062	**-0.138**	0.046
K	**-0.484**	**0.132**	**-0.181**	**-0.112**	0.038	**-0.212**	-0.038
Na	-0.043	0.045	**-0.155**	-0.043	0.079	**-0.175**	0.038
Mg	-0.015	0.020	**-0.104**	-0.019	0.002	-0.005	0.082
Ca	-0.027	0.026	-0.034	**-0.144**	0.036	-0.031	0.060

and seawater salinity, and the third is difficult to interpret. In particular, according to the model, seawater salinity acts at a greater range than carbonate dissolution. This is consistent with the fact that the "footprints" of seawater salinity in percolating groundwater (observed at the regional scale) are explained by the proximity of falling rain to the Mediterranean Sea, while increased carbonate dissolution is affected by factors acting at shorter ranges, such as the influence of sewage (ammonia oxidation) and vegetation (acid root exudates and decomposition of organics).
For the alternative methods, the synthesis algorithm failed to converge. There were no numerical problems with the first two alternatives, but the components were very difficult to interpret. Most of them were contrasts such as (HCO_3 - Ca), or combinations such as (HCO_3 + Cl + Ca) which had no physical intepretation in the context of the problem.

5 Discussion

The LMC model of (3), though not giving the factors z_{jk} in explicit form, is nonetheless plausible. The alternative methods give explicit expressions for factors which approximate special cases of the LMC: for example, the synthesis corresponds to a LMC with m sill matrices of rank one. For our data (and possibly for other problems as well), it may be that this specialized model is not appropriate, and thus the algorithm defines artificial combinations of the variables in order to fit the model to the data.

References

Bailey, T.C., and Krzanovski, W.J. (2000). Extensions to spatial factor methods with an illustration in geochemistry. *Mathematical Geology*, **32**,

657-682.

Clarkson, D.B. (1988a). Remark AS R74. A least squares version of algorithm AS 211: the F-G diagonalization algorithm. *Applied Statistics*, **37**, 317-321.

Clarkson, D.B. (1988b). Remark AS R71. A remark on algorithm AS 211: the F-G diagonalization algorithm. *Statistical algorithms, Applied Statistics*, **37**, 147-151.

Goovaerts, P. (1994). On a controversial method for modeling a coregionalization. *Mathematical Geology*, **26**, 197-204.

Goulard, M., and Voltz, M. (1992). Linear coregionalization model: Tools for estimation and choice of cross-varioram matrix. *Mathematical Geology*, **24**, 269-286.

Journel, A.G., and Huijbregts, Ch.J. (1978). *Mining geostatistics*. New York: Academic Press.

Krzanovski, W.J., and Bailey, T.C. (2007). Extraction of spatial features using factor methods illustrated on stream sediment data. *Mathematical Geology*, **39**, 69-85.

Maestre, F.T., Rodriguez, F., Bautista, S., Cortina, J., and Bellot, J. (2005). Spatial associations and patterns of perennial vegetation in a semiarid steppe: a multivariate geostatistics approach. *Plant Ecology*, **179**, 133-147.

Oman, S.D, and Vakulenko-Lagun, B. (2009). Estimation of sill matrices in the linear model of coregionalization. *Mathematical Geosciences*, **41**, 15-27.

Wackernagel, H. (1989). Description of a computer program for analyzing multivariate spatially distributed data. *Computers and Geosciences*, **15**, 593-598.

Wackernagel, H. (1994). Cokriging versus kriging in regionalized multivariate data analysis. *Geoderma*, **62**, 83-92.

Webster, R., Atteia, O., and Dubois, J.P. (1994). Coregionalization of trace metals in the soil in the Swiss Jura. *European Journal of Soil Science*, **45**, 205-218.

Spatial-Temporal Modelling of Extreme Rainfall

Mark Palmer[1], Eddy Campbell[1], Aloke Phatak[1], Bryson Bates[2]

[1] CSIRO Mathematics, Informatics and Statistics, PMB 5, Wembley, Western Australia 6913, Australia.
[2] CSIRO Marine and Atmospheric Research, PMB 5, Wembley, Western Australia 6913, Australia

Abstract: Extreme rainfall over two regions of Australia, the SW of Western Australia and the Sydney region of NSW, covering approximately the last fifty years, has been modelled using a Bayesian Hierarchical Model. A convolution kernel approach is used to derive Gaussian processes to model the spatial variability of the parameters of the Generalised Extreme Value distribution describing rainfall extremes. This is a flexible approach accommodating rainfall measured over different durations (from sub to super daily) and allowing for the possibility of linking the extremes to external drivers.

Keywords: convolution, Generalized Extreme Value distribution, spatial-temporal

1 Introduction

Changes in rainfall patterns are of great concern in Australia. Two areas of concern (Figure 1) have been studied: the western region had been considered one of the most reliable rainfall areas for growing wheat in Australia and the other is centred on a large urban population. We are studying changes over the period 1953-2003.

1.1 The data

Both daily and pluvio data (rainfall recorded over small time intervals, which one can aggregate) are available. It is important to utilise as much data as possible to develop accurate estimates. The data records for individual stations need not be necessarily complete for the region of study, either spatially or temporally.

1.2 The Spatial-Temporal Model

The parameters of the Generalised Extreme Value (GEV) distribution (Coles, 2001), known as the location, scale and shape parameters, are modelled as Gaussian processes which allow them to vary smoothly through space.

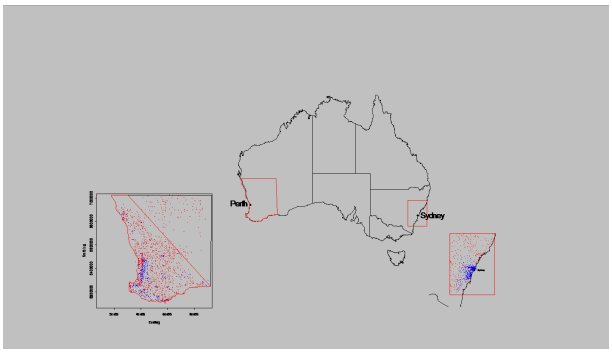


FIGURE 1. The study areas: The lower SW of Western Australia and mid New South Wales, showing daily rainfall stations (red) and pluvio stations (blue).

A convolution kernel approach (Higdon, 2002), based on convolving white noise with a suitable kernel, is used to derive these Gaussian processes. This is a very flexible approach to spatial modelling, able to cope with multivariate spatial correlation and non-stationarity, for example.
Covariates, such as ocean heat, are introduced to drive the parameters at a station through time, while other covariates such as height above sea level and distance from the coast model spatial trends in the parameters. Development of covariate selection procedures is currently being pursued. This results in a Bayesian hierarchical model, Figure 2, when priors for the parameters in the model are introduced. MCMC techniques (Smith et al. , 1993) are used to estimate parameters and derive measures of variability.

2 Conclusion

This model is very flexible in allowing us to incorporate an increased amount of data, and also for the incorporation of covariates. These covariates can be derived from other computer models and this then allows us to predict changes in extreme rainfall patterns under changing climate patterns. Figure 3 shows an increase in return levels close to the coast and a decrease further inland when driven by ocean heat content.
From the model we can derive Intensity-Duration-Frequency curves, Figure 4, which are important tools in deriving engineering specifications for dams, culverts road works etc. Importantly, because of the Bayesian approach, it is also possible to derive measures of variability associated with these curves. Extensions of this model to estimate Depth area curves are being investigated.

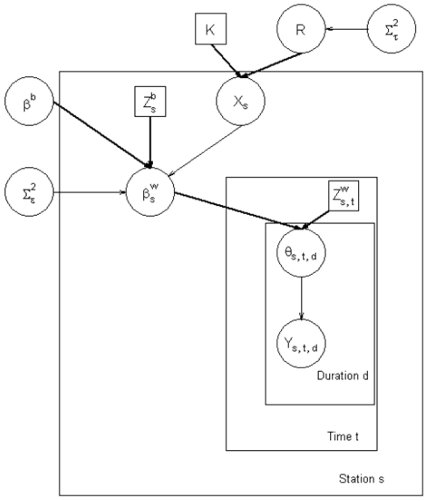


FIGURE 2. Di-graph representation of the Spatial-Temporal model of extreme rainfall.

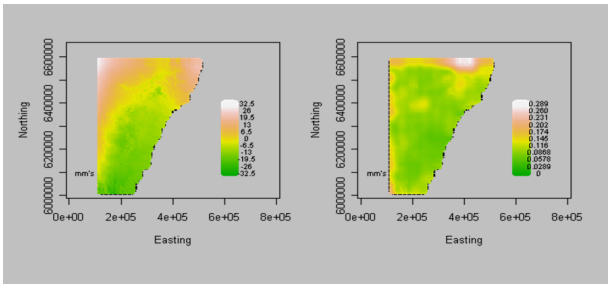


FIGURE 3. Differenced return levels surfaces (2003 1953) for a fifty year return period (a), and associated standard errors (b), driven by ocean heat.

Acknowledgments: Organizations involved in these projects include: the Australian Greenhouse Office, the Upper Parramatta River Catchment Trust, Sydney Water, Sydney Metropolitan Catchment Management Authority, the Hunter-Central Rivers Catchment Management Authority, the Southern Rivers Catchment Management Authority and the Indian Ocean

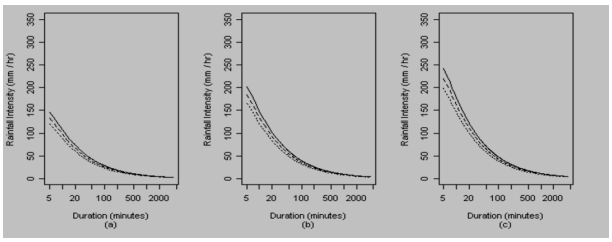


FIGURE 4. Estimated average IDF curves, for return periods of (a) 5 years, (b) 20 years and (c) 50 years. Each figure shows the IDF curves calculated using an ocean heat anomaly of -2.5, 0.0 and 2.5 respectively. Increasing values of the ocean heat anomaly lead to lower IDF curves within each figure.

Climate Initiative. This work has been undertaken as part of the Australian Climate Change Science Program, funded jointly by the Department of Climate Change, the Bureau of Meteorology and CSIRO.'

References

Coles, S. (2001). *An Introduction to Statistical Modelling of Extreme Values*. New York: Springer.

Higdon, D. (2002). Space and space-time modelling using process convolutions. In: *Quantitative methods for current environmental issues*, C. Anderson, V. Barnett, P. Chatwin, and A. El-Shararawi, Eds., Springer-Verlag, pp 37-56.

Smith, A. F. M., and G.O. Roberts (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J.R. Statist Soc. B*, **6**, 55, 3-23.

Restricted Maximum Likelihood Estimation in Joint Mean-Covariance Models

Georgios Papageorgiou[1]

[1] School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway

Abstract: The class of joint mean-covariance models uses the modified Cholesky decomposition of the within subject covariance matrix in order to arrive to an unconstrained, statistically meaningful reparameterisation. Furthermore, it is well known that, in all classes of models, maximum likelihood estimation provides overoptimistic variance components estimates as it does not take into account the loss of degrees of freedom that results from estimating the mean structure. Here, we propose adjustments to the estimating equations in order to alleviate the problem of inefficient estimation and downward bias in the class of joint mean-covariance models.

Keywords: adjusted profile likelihood; Cholesky decomposition; longitudinal data.

1 Joint mean-covariance models and maximum likelihood estimation

In longitudinal studies, repeated measures on the same subjects are taken over time, and these typically are correlated. Several approaches have been proposed for taking this correlation into account, such as the linear mixed models (Laird and Ware, 1982), the generalized linear mixed models (Breslow and Clayton, 1993) and the joint mean-covariance models (Pourahmadi, 1999, 2000).

Our focus here is on the joint mean-covariance models which model both the mean and covariance structures in terms of covariates. The underlying assumption in these models is that the responses $\mathbf{Y}_i$ are distributed according to $\mathbf{Y}_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, and the goal is to find parsimonious models for both the $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$. The main difficulty in modeling covariance structures is the constraint that the covariance matrix is positive definite. The joint mean-covariance models, by utilizing the modified Cholesky decomposition of the within-subject covariance matrix, result in an unconstrained and statistically meaningful reparameterisation.

Specifically, the covariance matrix $\boldsymbol{\Sigma}$ of a random vector $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ can be diagonalized using $\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}^T = \mathbf{D}$, where $\mathbf{T}$ is a unique lower triangular matrix with ones on its diagonal and $\mathbf{D}$ is a unique diagonal matrix with

positive entries. The new covariance parameters, that is the non-redundant parameters in $\mathbf{T}$ and $\mathbf{D}$, for a time ordered random vector, are interpreted as the prediction coefficients and prediction error variances that are associated with the prediction of a response based on its predecessors. In particular, with $\mu_t = E(Y_t)$, the linear least squares predictor of a response y_t based on $y_{t-1}, \ldots, y_1$ is $\hat{y}_t = \mu_t + \sum_{j=1}^{t-1} \phi_{t,j}(y_j - \mu_j)$ and the prediction error variance is $\sigma_t = \text{var}(y_t - \hat{y}_t)$. The below diagonal entries of $\mathbf{T}$ are the negatives of $\phi_{t,j}$, while the diagonal entries of $\mathbf{D}$ are the prediction error variances, σ_t. This makes clear that there are no constraints on the parameters $\phi_{t,j}$ and $\log(\sigma_t)$ and thus they can be modeled using covariates, just as it is done for the mean response in linear models.
For $t = 1, \ldots, n$ and $j = 1, \ldots, t-1$, these models take the form:

$$\mu_t = \mathbf{x}_t^T \boldsymbol{\beta}, \log(\sigma_t) = \mathbf{z}_t^T \boldsymbol{\lambda}, \text{and } \phi_{t,j} = \mathbf{z}_{t,j}^T \boldsymbol{\gamma},$$

where $\mathbf{x}_t$, $\mathbf{z}_t$ and $\mathbf{z}_{t,j}$ are $p \times 1$, $q \times 1$ and $d \times 1$ design vectors, and $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ are the mean, variance and correlation parameters respectively.
Maximum likelihood (ML) estimation is achieved by a Newton-Raphson or a Fisher scoring algorithm (Pourahmadi, 1999, 2000). Using either of these to algorithms, the objective is to find solutions to the score equations:

$$\mathbf{U}_1(\boldsymbol{\beta}; \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \mathbf{0}, \mathbf{U}_2(\boldsymbol{\lambda}; \boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathbf{0}, \mathbf{U}_3(\boldsymbol{\gamma}; \boldsymbol{\beta}, \boldsymbol{\lambda}) = \mathbf{0},$$

which provide us the ML estimates $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\gamma}}$ respectively.

2 Adjustments of the estimating equations

Some of the basic properties of the estimating equations, $\mathbf{U}_i$, $i = 1, 2, 3$, are that they are unbiased, that is their expectation is zero, and information unbiased, that is their variance is minus their expected derivative matrix, where all expectations are calculated at the true parameter value.
Maximum likelihood estimation proceeds by finding the solution, $\hat{\boldsymbol{\beta}}$, to $\mathbf{U}_1$ and replacing the unknown $\boldsymbol{\beta}$ in $\mathbf{U}_i$, $i = 2, 3$ by $\hat{\boldsymbol{\beta}}$. The resulting profile estimating equations, $\mathbf{U}_2(\boldsymbol{\lambda}; \hat{\boldsymbol{\beta}}, \boldsymbol{\gamma})$ and $\mathbf{U}_3(\boldsymbol{\gamma}; \hat{\boldsymbol{\beta}}, \boldsymbol{\lambda})$, however, are no longer unbiased nor information unbiased and thus the resulting estimates can be inefficient, especially when the dimension of $\hat{\boldsymbol{\beta}}$ is high.
Here, for the joint mean-covariance models, we propose 'degree of freedom' adjustments, in the spirit of McCullagh and Tibshirani (1990), in order to alleviate the problem of downward estimation of the covariance parameters. Specifically, we adjust the profile estimating equations so that they have mean zero. Further adjusting these equations in order to make them information unbiased does not provide any benefit in the context of mean-covariance models and so it is not further pursued. The proposed adjustments are described in more detail in the following paragraphs.

As it was shown by Pourahmadi (2000), for the case of balanced longitudinal data, the estimating equation $\mathbf{U}_2$ and $\mathbf{U}_3$ are given by:

$$\mathbf{U}_2 = \frac{1}{2}\mathbf{Z}^T\left(\mathbf{D}^{-1}\mathbf{R} - m\mathbf{1}_n\right) \text{ and } \mathbf{U}_3 = \sum_{i=1}^{m}\mathbf{Z}_{\boldsymbol{\beta}}(i)^T\mathbf{D}^{-1}\left(\mathbf{r}_i - \mathbf{Z}_{\boldsymbol{\beta}}(i)\boldsymbol{\gamma}\right) \quad (1)$$

where $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^T$, $\mathbf{r}_i(\boldsymbol{\beta}) = \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}$, $\mathbf{R}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathrm{RSS}_1, \ldots, \mathrm{RSS}_n)^T$ where $\mathrm{RSS}_t = \sum_{i=1}^{m}(r_{it} - \hat{r}_{it})^2$ with $\hat{r}_{it}$ being the predictor of r_{it} based on its predecessors $r_{i,t-1}, \ldots, r_{i1}$, and, finally, $\mathbf{Z}_{\boldsymbol{\beta}}(i) = (\mathbf{z}(i,1), \ldots, \mathbf{z}(i,n))^T$ with $\mathbf{z}(i,t) = \sum_{j=1}^{t-1} r_{ij}\mathbf{z}_{tj}$, $t = 1, \ldots, n$.

It can be shown that $E\{\mathbf{U}_2(\boldsymbol{\lambda}; \hat{\boldsymbol{\beta}}, \boldsymbol{\gamma})\} = -2^{-1}\mathbf{Z}^T\mathbf{D}^{-1}\mathbf{b}$, where $\mathbf{b} \equiv \mathbf{b}(\boldsymbol{\lambda}, \boldsymbol{\gamma})$ is the vector of the diagonal elements of the matrix $\mathbf{T}(\boldsymbol{\gamma})\mathbf{B}(\boldsymbol{\lambda}, \boldsymbol{\gamma})\mathbf{T}(\boldsymbol{\gamma})^T$, where $\mathbf{B} \equiv \mathbf{B}(\boldsymbol{\lambda}, \boldsymbol{\gamma}) = \sum_{i=1}^{m}\mathbf{X}_i(\sum_{k=1}^{m}\mathbf{X}_k^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_k)^{-1}\mathbf{X}_i^T$. It can also be shown that $E\{\mathbf{U}_3(\boldsymbol{\gamma}; \hat{\boldsymbol{\beta}}, \boldsymbol{\lambda})\} = -\boldsymbol{\Delta}(\boldsymbol{\lambda}, \boldsymbol{\gamma})^T\mathbf{D}^{-1}\mathbf{1}_n$, where $\boldsymbol{\Delta} \equiv \boldsymbol{\Delta}(\boldsymbol{\lambda}, \boldsymbol{\gamma}) = (\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_n)^T$ and, for $t = 1 \ldots, n$, $\boldsymbol{\Delta}_t = \sum_{j=1}^{t-1} a_{tj}\mathbf{Z}_{\mathbf{tj}}$ where a_{tj} is the (t, j) element of the product matrix $\mathbf{TB}$.

Subtracting the respective biases of equations $\mathbf{U}_2$ and $\mathbf{U}_3$ we obtain the bias corrected estimating equations

$$\tilde{\mathbf{U}}_2(\boldsymbol{\lambda}; \boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2}\mathbf{Z}^T\left\{\mathbf{D}(\boldsymbol{\lambda})^{-1}\left[\mathbf{R}(\boldsymbol{\beta}, \boldsymbol{\gamma}) + \mathbf{b}(\boldsymbol{\lambda}, \boldsymbol{\gamma})\right] - m\mathbf{1}_n\right\} \quad (2)$$

$$\tilde{\mathbf{U}}_3(\boldsymbol{\gamma}; \boldsymbol{\beta}, \boldsymbol{\lambda}) = \sum_{i=1}^{m}\mathbf{Z}_{\boldsymbol{\beta}}(i)^T\mathbf{D}(\boldsymbol{\lambda})^{-1}\left(\mathbf{r}_i - \mathbf{Z}_{\boldsymbol{\beta}}(i)\boldsymbol{\gamma}\right) + \boldsymbol{\Delta}^T\mathbf{D}(\boldsymbol{\lambda})^{-1}\mathbf{1}_n \quad (3)$$

in which the unknown parameter $\boldsymbol{\beta}$ can be replaced by its ML estimate.

A further adjustment, however, is needed in order to be able to replace $\boldsymbol{\gamma}$ by $\hat{\boldsymbol{\gamma}}$ in $\tilde{\mathbf{U}}_2$ and $\boldsymbol{\lambda}$ by $\hat{\boldsymbol{\lambda}}$ in $\tilde{\mathbf{U}}_3$. Due to the space constraint, here we only describe the proposed adjustment of $\tilde{\mathbf{U}}_2$. The adjustment of $\tilde{\mathbf{U}}_3$ can be done in a similar way.

We first observe that $\tilde{\mathbf{U}}_2$ depends on $\boldsymbol{\gamma}$ through $\mathbf{G}(\boldsymbol{\gamma}) \equiv \mathbf{R}(\boldsymbol{\beta}, \boldsymbol{\gamma}) + \mathbf{b}(\boldsymbol{\lambda}, \boldsymbol{\gamma})$. Now, as was also done by Ghosh and Maiti (2004), by a two step Taylor expansion of $\mathbf{G}_k(\boldsymbol{\gamma})$ around $\hat{\boldsymbol{\gamma}}$

$$\mathbf{G}_k(\hat{\boldsymbol{\gamma}}) + \frac{\partial\mathbf{G}_k(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}^T(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}) + \frac{1}{2}\mathrm{tr}\{\frac{\partial^2\mathbf{G}_k(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T}(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}})(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}})^T\} \quad (4)$$

where $\mathbf{G}_k$ if the kth element of $\mathbf{G}$. In (4) the unknown quantities $\boldsymbol{\gamma}$-$\hat{\boldsymbol{\gamma}}$ and $(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}})(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}})^T$ are replaced by their expectations. It is a standard result that $E\{(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}})(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}})^T\} = \boldsymbol{I}_{\boldsymbol{\gamma}}{}^{-1}$, where $\boldsymbol{I}_{\boldsymbol{\gamma}} = E\{-\partial\mathbf{U}_3/\partial\boldsymbol{\gamma}\}$. The formula for $\boldsymbol{I}_{\boldsymbol{\gamma}}$ has been provided by Pourahmadi (2000). Finding $E\{\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}\}$ is more challenging and we defer it to the Appendix where we also provide some formulas necessary for calculation of the first and second order derivatives of $\mathbf{G}(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$.

Having obtained expressions for all the quantities in (4), we replace $\mathbf{G}(\boldsymbol{\gamma})$ in (2) by a vector that has kth element, $k = 1, \ldots, n$, equal to

$$\mathbf{G}_k(\hat{\boldsymbol{\gamma}}) + \frac{\partial \mathbf{G}_k(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}^T E\{\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}\} + \frac{1}{2}\mathrm{tr}\{\frac{\partial^2 \mathbf{G}_k(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T} \boldsymbol{I}_{\boldsymbol{\gamma}}{}^{-1}\}, \tag{5}$$

where all expressions evaluated at $\boldsymbol{\gamma}$=$\hat{\boldsymbol{\gamma}}$. We denote the resulting equation by $\mathbf{U}_2^*$. The similarly corrected estimating equation of $\boldsymbol{\gamma}$ is denoted by $\mathbf{U}_3^*$. We set $\mathbf{U}^* = \{(\mathbf{U}_2^*)^T, (\mathbf{U}_3^*)^T\}^T$.

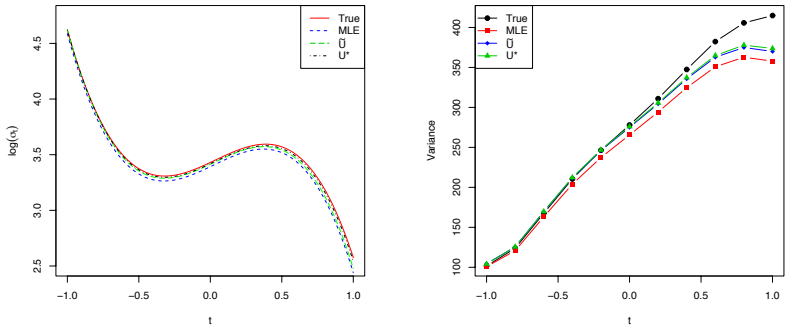
3 Application to the cattle data and simulation results

We apply the proposed estimating procedure to Kenward's (1987) cattle data. In this balanced longitudinal study, 60 cattle received one of two treatments for intestinal parasites. Measurements on the weights of the cattle were taken $n = 11$ times during a time period of 133 days.
Following Pourahmadi (1999, 2000), we analyze the data of the $m = 30$ cattle that received treatment A. However, we choose to fit the best fitting model according to BIC (Pan and Mackenzie (2003)). That is, we assume a polynomial of degree eight for the mean structure, a third degree polynomial model for the innovation variances, σ_t, and a fourth degree polynomial for the autoregressive coefficients, $\phi_{t,j}$:

$$\log(\sigma_t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 \tag{6}$$

$$\phi_{t,j} = \delta_0 + \delta_1(t-j) + \delta_2(t-j)^2 + \delta_3(t-j)^3 + \delta_4(t-j)^4. \tag{7}$$

Based on this model, we reanalyze the cattle data and obtain ML estimates for the model parameters. Subsequently, we use these values as the true model values in a simulation study. At each of the 250 runs of the simulation study, we obtain ML estimates, and estimates based on the bias corrected estimating equations $\tilde{\mathbf{U}}$ and $\mathbf{U}^*$. Part of our results are shown in Figure 1. First, from Figure 1(a) becomes evident that, on average, the ML estimates of the logarithms of the innovation variances, $\log(\sigma_t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3$, underestimate the true values of these. The estimates obtained from the bias corrected estimating equations, $\tilde{\mathbf{U}}$, also underestimate the true values of the logarithms of the innovation variances. However, in the estimates obtained from $\tilde{\mathbf{U}}$, the problem is alleviated. A further slight improvement in the estimation of the innovation variances is achieved by $\mathbf{U}^*$. Further, in Figure 1(b) we plot the true and estimated values of the diagonal elements of the covariance matrix $\boldsymbol{\Sigma}$, which we reconstruct from the innovation variances and autoregressive coefficients. Clearly, due to the underestimation in the innovation variances, the variances are also underestimated. However, we see that, on average, the corrected equations provide estimates that are closer to the true value than the ML estimates and that $\mathbf{U}^*$ slightly improves $\tilde{\mathbf{U}}$.



(a) (b)

FIGURE 1. Simulation results based on 250 replications: (a) True and average estimated logarithms of the innovation variances, (b) True and average estimated diagonal elements of the marginal covariance matrix. Estimation was carried out using the ML approach and the bias corrected estimating equations: $\tilde{\mathbf{U}}$ and $\mathbf{U}^*$.

4 Appendix

In order to find an expression for $E\{\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}\}$ we start by defining $L \equiv L(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma})$ to be the log-likelihood. Further, following Cox and Snell (1968), let $J_{r,sc} = \text{Cov}\left(\frac{\partial L}{\partial \boldsymbol{\gamma}_s}, \frac{\partial^2 L}{\partial \boldsymbol{\gamma}_r \partial \boldsymbol{\gamma}_c}\right)$, for $r, s, c = 1, \ldots, d$. Now let $\mathbf{J}_r$ be the $d \times d$ matrix that has (s, c) element equal to $J_{r,sc}$. We now write $\mathbf{A}^T = (\text{tr}(\boldsymbol{I}_{\boldsymbol{\gamma}}{}^{-1}\mathbf{J}_1), \ldots, \text{tr}(\boldsymbol{I}_{\boldsymbol{\gamma}}{}^{-1}\mathbf{J}_d))$. By the Cox-Snell formula we have that $E\{\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}\} = -\boldsymbol{I}_{\boldsymbol{\gamma}}{}^{-1}\mathbf{A}$.
The first and second order derivatives of $\mathbf{R}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ are found using the following formulas

$$\partial \text{RSS}_t / \partial \boldsymbol{\gamma} = -2 \sum_{i=1}^{m} r_{it} \mathbf{z}(i,t) + 2 \sum_{i=1}^{m} \mathbf{z}(i,t)\mathbf{z}(i,t)^T \boldsymbol{\gamma}$$

$$\partial^2 \text{RSS}_t / (\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T) = 2 \sum_{i=1}^{m} \mathbf{z}(i,t)\mathbf{z}(i,t)^T.$$

The first order derivatives of $\mathbf{b}(\boldsymbol{\lambda}, \boldsymbol{\gamma})$ are calculated based on

$$\partial \mathbf{B} / \partial \boldsymbol{\gamma}_k = \sum_{i=1}^{m} \mathbf{X}_i (\sum_{k=1}^{m} \mathbf{X}_k^T \boldsymbol{\Sigma}^{-1} \mathbf{X}_k)^{-1} \mathbf{V} (\sum_{k=1}^{m} \mathbf{X}_k^T \boldsymbol{\Sigma}^{-1} \mathbf{X}_k)^{-1} \mathbf{X}_i^T, \qquad (8)$$

where $\mathbf{V} = \sum_{k=1}^{m} \mathbf{X}_k^T \{-\mathbf{T}^T \mathbf{D}^{-1} \frac{\partial \mathbf{T}}{\partial \boldsymbol{\gamma}_k} - \frac{\partial \mathbf{T}^T}{\partial \boldsymbol{\gamma}_k} \mathbf{D}^{-1} \mathbf{T}\} \mathbf{X}_k$. Further, calculation

of $\partial^2\mathbf{B}/(\partial\boldsymbol{\gamma}_k\partial\boldsymbol{\gamma}_j)$ uses (8) and

$$\frac{\partial}{\partial\boldsymbol{\gamma}_j}(\sum_{k=1}^{m}\mathbf{X}_k^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_k)^{-1} = (\sum_{k=1}^{m}\mathbf{X}_k^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_k)^{-1}\mathbf{P}(\sum_{k=1}^{m}\mathbf{X}_k^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_k)^{-1}, \quad (9)$$

where $\mathbf{P} = \sum_{k=1}^{m}\mathbf{X}_k^T\left\{\mathbf{T}^T\mathbf{D}^{-1}\frac{\partial\mathbf{T}}{\partial\boldsymbol{\gamma}_j} - \frac{\partial\mathbf{T}^T}{\partial\boldsymbol{\gamma}_j}\mathbf{D}^{-1}\mathbf{T}\right\}\mathbf{X}_k$. Finally, the following is also needed

$$\frac{\partial}{\partial\boldsymbol{\gamma}_j}\left\{\mathbf{T}^T\mathbf{D}^{-1}\frac{\partial\mathbf{T}}{\partial\boldsymbol{\gamma}_k}\right\} = \frac{\partial\mathbf{T}^T}{\partial\boldsymbol{\gamma}_j}\mathbf{D}^{-1}\frac{\partial\mathbf{T}}{\partial\boldsymbol{\gamma}_k}.$$

Acknowledgments: This research was supported by Science Foundation Ireland Research Frontiers grant 07/RFP/MATF448.

References

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

Cox, D.R., and Snell, E.J. (1968). A general definition of residuals (with Discussion). *Journal of the Royal Statistical Society Series B*, **30**, 248-275.

Ghosh, M., and Maiti, T. (2004). Small-area estimation based on natural exponential family quadratic variance function models and survey weights. *Biometrika*, **91**, 95-112.

Kenward, M.G. (1987). A Method for comparing profiles of repeated measurements. *Applied Statistics*, **36**, 296-308.

McCullagh, P., and Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society, Series B*, **52**, 325-344.

Laird, N.M., and Ware, J.J. (1982). Random effects models for longitudinal data. *Biometrics*, **38**, 973-979.

Pan, J., and Mackenzie, G. (2003). On modelling mean-covariance structures in longitudinal studies. *Biometrika*, **90**, 239-244.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, **86**, 677-690.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, **87**, 425-435.

Fitting Global Cross-Ratio Models for Longitudinal data by Minimum phi-Divergence

M.C. Pardo[1], R. Alonso[1]

[1] Department of Statistics and O.R. (I), Complutense University of Madrid, Spain

Abstract: In this paper, we focus on repeated measurement problems, comprising an interesting research area in statistics. We study longitudinal data which arise when outcomes are observed repeatedly on each experimental subject at several points. We focus on a marginal approach for this type of data with lack of independence among the observations. We propose an alternative estimation based on divergence measures to a full likelihood method for a family of statistical models proposed by Dale for bivariate, discrete response. Finally, a clinical data set will be used to illustrate the new procedure.

Keywords: Longitudinal Data, odds ratios, ordinal data, minimum $\phi-$divergence estimator.

1 Introduction

Longitudinal data are frequent in biological and medical experimental research, where individuals are observed over time (Agresti, 2002; Diggle et al., 2002). The longitudinal studies require special statistical methods because the set of observations taken on one unit are usually intercorrelated. An important consideration in the statistical modelling of correlated data concerns the type of outcome. Methods for continuous data are doubtless the best developed and the linear mixed model (Laird and Ware (1982), Verbeke and Molenberghs (1997,2000), has played a prominent role in extending the generalized linear model to handle correlated continuous data. Owing to the elegant properties of the multivariate normal distribution, its theory and implementation are greatly simplified.
When the outcome variable is discrete (e.g. counts) or categorical (nominal or ordinal data), a first issue arises which is the lack of a discrete analogue to the multivariate normal distribution. An approach leads to a class of regression models known as marginal or population-averaged models. In particular, we focus on a marginal likelihood model, the Global Cross-Ratio Model, introduced by Dale (1986) in which the within-cluster association can be parametrized in terms of marginal odds ratios. This model is appropriate for bivariate, discrete, ordered responses.

Section 2 is devoted to introduce the Global Cross-Ratio Models. The fitting problem of the model is considered in Section 3. The results are illustrated in a numerical example in Section 4.

2 Global Cross-Ratio Models

Let $\mathbf{Z}(\mathbf{x}) = (Z_1(\mathbf{x}), Z_2(\mathbf{x}))$ be a discrete random vector, dependent on a fixed p-dimensional covariate $\mathbf{x}$. Suppose that $Z_1(\mathbf{x})$ takes the value $1, ..., r$ according to the cumulative probabilities $\eta_{i\mathbf{x}} = \Pr(Z_1(\mathbf{x}) \leq i), i = 1, ..., r$, and similarly that the cumulative probability function for $Z_2(\mathbf{x})$ is $\xi_{j\mathbf{x}} = \Pr(Z_2(\mathbf{x}) \leq j), j = 1, ..., c$.
The cumulative probability function $\mathbf{Z}(\mathbf{x})$ is

$$F_{ij}(\mathbf{x}; \theta_{\mathbf{x}}) = \Pr(Z_1(\mathbf{x}) \leq i, Z_2(\mathbf{x}) \leq j),$$

with $\theta_{\mathbf{x}} = (\eta_{\mathbf{x}}, \xi_{\mathbf{x}}, \psi_{\mathbf{x}})$, where $\eta_{\mathbf{x}} = (\eta_{1\mathbf{x}}, ..., \eta_{r-1\mathbf{x}})^T, \xi_{\mathbf{x}} = (\xi_{1\mathbf{x}}, ..., \xi_{c-1\mathbf{x}})^T$ and $\psi_{\mathbf{x}}$ is a matrix of global ratios to be defined. Now divide the space of possible values of $\mathbf{Z}$ into four quadrants,

$$\{Z_1(\mathbf{x}) \leq i, Z_2(\mathbf{x}) \leq j\}, \{Z_1(\mathbf{x}) \leq i, Z_2(\mathbf{x}) > j\},$$
$$\{Z_1(\mathbf{x}) > i, Z_2(\mathbf{x}) \leq j\}, \{Z_1(\mathbf{x}) > i, Z_2(\mathbf{x}) > j\}.$$

This double dichotomy is said to occur at bivariate 'cutpoint' (i, j). The $r \times c$ table of probabilities $\pi_{ij/\mathbf{x}} = \Pr(Z_1(\mathbf{x}) = i, Z_2(\mathbf{x}) = j)$ is thus collapsed at cutpoint (i, j) to a 2×2 table. Define the 'global cross-ratio' (GCR) at cutpoint (i, j) to be

$$\psi_{ij\mathbf{x}} = \frac{\Pr(Z_1(\mathbf{x}) \leq i, Z_2(\mathbf{x}) \leq j) \Pr(Z_1(\mathbf{x}) > i, Z_2(\mathbf{x}) > j)}{\Pr(Z_1(\mathbf{x}) > i, Z_2(\mathbf{x}) \leq j) \Pr(Z_1(\mathbf{x}) \leq i, Z_2(\mathbf{x}) > j)}.$$

A general form to modelize $\log(\psi_{ij\mathbf{x}})$ is

$$\log(\psi_{ij\mathbf{x}}) = \Delta + \alpha_{ia} + \beta_{ja} + \delta_{ij} - \gamma^T \mathbf{x}, \tag{1}$$

$i = 1, ..., r-1, j = 1, ...c-1, a =$association, with simple uniqueness constrains such as $\alpha_{r-1,a} = \beta_{c-1,a} = 0; \delta_{i,c-1} = 0, i = 1, ..., r-1; \delta_{r-1,j} = 0, j = 1, ..., c-1$.
The cumulative marginal probabilities are fitted using the generalized linear models:

$$\begin{aligned} g_1(\eta_{i\mathbf{x}}) &= \alpha_{1i} - \mathbf{x}^T \beta_1, i = 1, ..., r-1; \\ g_2(\xi_{j\mathbf{x}}) &= \alpha_{2j} - \mathbf{x}^T \beta_2, j = 1, ..., c-1. \end{aligned} \tag{2}$$

3 Fitting the Model

A straightforward method of estimation with desirable asymptotic properties in this situation is maximum likelihood (Cox and Hinkley, (1974)).
Consider n_k independent observations distributed according to the GCR family of models with covariate $\mathbf{x}_k$, linear logistic margins obtained from (2) for $g_q(s) = \log(s/(1-s)), q = 1, 2$, and parameters as in (1).
Let Y_{ijk} be the number of observations in cell (i, j), with $\mathbf{Y}_k$ denoting the whole $r \times c$ table. Then the distribution of $\mathbf{Y}_k$ is multinomial on n_k observations and $r \times c$ cells. We will fit this GCR model to a sequence of m independent tables $\mathbf{Y}_1, ..., \mathbf{Y}_m$ dependent on a matrix of covariates $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_m)$. The log-likelihood function for $\mathbf{Y}_1, ..., \mathbf{Y}_m$, up to an additive function not dependent on θ, is

$$l(\theta; \mathbf{y}_1, ..., \mathbf{y}_m, \mathbf{X}) = \sum_{k=1}^{m} \sum_{i=1}^{r} \sum_{j=1}^{c} y_{ijk} \log(\pi_{ij}(\mathbf{x}_k; \theta))$$

where

$$\begin{aligned} \pi_{ij}(\mathbf{x}_k; \theta) &= \Pr(Z_1(\mathbf{x}_k) = i, Z_2(\mathbf{x}_k) = j \mid \theta), \\ \theta &= (\alpha_1, \beta_1, \alpha_2, \beta_2, \Delta, \alpha_a, \beta_a, \delta, \gamma), \end{aligned}$$

and the components of θ are defined in (2) and (1).
Now to obtain the maximum likelihood estimator (MLE) l must be maximized. However, it can be proved that maximize l is equivalent to minimize the Kullback divergence between the probability vectors $\widehat{\mathbf{p}}$ and $\mathbf{p}(\theta)$ given by

$$D_{Kullback}(\widehat{\mathbf{p}}, \mathbf{p}(\theta)) = \sum_{k=1}^{m} \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{y_{ijk}}{n} \log \frac{\frac{y_{ijk}}{n}}{\pi_{ij}(\mathbf{x}_k; \theta) \frac{n_k}{n}}$$

being $\widehat{\mathbf{p}} = \left(\frac{y_{111}}{n}, ..., \frac{y_{mrc}}{n}\right)$, $\mathbf{p}(\theta) = \left(\pi_{11}(\mathbf{x}_1; \theta) \frac{n_1}{n}, ..., \pi_{rc}(\mathbf{x}_m; \theta) \frac{n_m}{n}\right)$ and $n = \sum_{k=1}^{m} n_k$.
Ought to this equivalence the MLE can be alternatively defined as the vector $\widehat{\theta}$ that verifies

$$D_{Kullback}\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\widehat{\theta}\right)\right) = \min_{\theta} D_{Kullback}(\widehat{\mathbf{p}}, \mathbf{p}(\theta)). \tag{3}$$

From this alternative way to define the MLE, a natural generalization of it consist on substituting the Kullback divergence for every divergence measure. For example, we focus on the ϕ-divergence family which is defined between two probability vectors $\mathbf{p} = (p_1, ..., p_M)^T$ and $\mathbf{q} = (q_1, ..., q_M)^T$ as

$$D_{\phi}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{M} q_i \phi\left(\frac{p_i}{q_i}\right), \qquad \phi \in \Phi^*,$$

where Φ^* is a class of convex functions $\phi : [0, \infty) \mapsto R \cup \{\infty\}$ such as in $x = 1$, $\phi(1) = 0$, $\phi''(1) > 0$ and in $x = 0$, $0\phi(0/0) = 0$ and $0\phi(p/0) = p \lim_{u\to\infty} \phi(u)/u$. See Pardo (2006).
Thus the minimum ϕ-divergence estimator is defined as the vector $\widehat{\theta}_\phi$ that verifies

$$D_\phi\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\widehat{\theta}_\phi\right)\right) = \min_\theta D_\phi\left(\widehat{\mathbf{p}}, \mathbf{p}(\theta)\right). \tag{4}$$

For $\phi(x) = x \log x - x + 1$, we obtain the MLE proposed by Dale (1986). It can be proved that the minimum ϕ-divergence estimator as the MLE is consistent and asymptotically multivariate normal.
Therefore, we propose to fit the GCR models using the minimum ϕ-divergence estimators. To obtain these new estimators as well as to obtain the MLE is necessary to use a numerical approach since a close form is not available.

4 Numerical illustration

As an illustration of the new family of estimators we consider data on relationship between patients' postoperative pain level and medication requirements on surgery for duodenal ulcer which are studied in Dale (1986). Let $\mathbf{x}$ represent the type of operation the patient was allocated (4 operation types). Therefore $\mathbf{x}$ is a vector of zeros with 1 in position i for $i = 1, ..., 4$. Let $Z_1(\mathbf{x})$ denote his pain level (none/slight/moderate) and let $Z_2(\mathbf{x})$ stand for the frequency of medication (never/seldom/occasionally/regularly). We focus on the most appropriate model proposed in that paper

$$\log(\psi_{ij\mathbf{x}}) = \Delta + \beta_{ja} + \gamma_3 \tag{5}$$

assuming that the marginal parameters are independent of operations.
To choose the best ϕ-divergence measure in the large class of estimators (4), it is not possible in general, we must choice a parameter family of ϕ-divergences for it. We shall consider $\phi = \phi_{(\lambda)}$ defined by Cressie and Read (1984) as

$$\phi_{(\lambda)}(x) = \frac{1}{\lambda(\lambda+1)}\left(x^{\lambda+1} - (x-1)\lambda\right);\ \lambda \neq 0,\ \lambda > -1 \tag{6}$$

where

$$\phi_{(0)}(x) = \lim_{\lambda\longrightarrow 0} \phi_{(\lambda)}(x) = xlogx - x + 1.$$

Adopting the name of the divergence measure which is used to obtain the different estimators, we have for $\lambda \to 0$ the maximum likelihood estimator, for $\lambda = 1$ the minimum chi-squared estimator, for $\lambda = -1/2$ the minimum Matusita distance (or Hellinger distance) estimator and for $\lambda = 2/3$ the minimum Cressie-Read estimator. For $\lambda \leq -1$ the minimum $\phi_{(\lambda)}-$divergence estimator is not defined if there are one or more empty cells. For minimizing $D_{\phi_{(\lambda)}}\left(\widehat{\mathbf{p}}, \mathbf{p}(\theta)\right)$ with respect θ to obtain $\widehat{\theta}_{\phi_{(\lambda)}}$ it was

used the NLPTR subroutine from SAS. The gradient of $D_{\phi_{(\lambda)}}(\widehat{\mathbf{p}}, \mathbf{p}(\theta))$ with respect θ was calculated algebraically; it is provided by the authors by request.

Parameter estimates for the model (5) using the estimators $\widehat{\theta}_{\phi_{(-1/2)}}$, $\widehat{\theta}_{\phi_{(0)}}$, $\widehat{\theta}_{\phi_{(2/3)}}$ and $\widehat{\theta}_{\phi_{(1)}}$ are given in the below tables as well as their standard error (SE). The SE was obtained as the square of the diagonal of minus the inverse of the second partial derivations matrix obtained numerically.

Parameters	$\widehat{\theta}_{\phi_{(-1/2)}}$	SE		$\widehat{\theta}_{\phi_{(0)}}$	SE
Association:					
Δ	4.0070	0.3083		3.8333	0.3048
β_{1a}^{a}	−1.2529	0.2834		−1.0806	0.2793
β_{2a}^{a}	−0.8716	0.2598		−0.7872	0.2529
γ_3	0.7644	0.3491		0.7617	0.3511
Marginal:					
α_{11}	1.1060	0.0717		1.0858	0.0713
α_{12}	2.1990	0.1028		2.1677	0.1014
α_{21}	1.4719	0.0806		1.4350	0.0795
α_{22}	1.9309	0.0940		1.9060	0.0932
α_{23}	2.7546	0.1310		2.7289	0.1303

Parameters	$\widehat{\theta}_{\phi_{(2/3)}}$	SE		$\widehat{\theta}_{\phi_{(1)}}$	SE
Association:					
Δ	3.7165	0.2989		3.6744	0.2958
β_{1a}^{a}	−0.9600	0.2770		−0.9148	0.2762
β_{2a}^{a}	−0.7264	0.2483		−0.7041	0.2462
γ_3	0.7747	0.3566		0.7848	0.3594
Marginal:					
α_{11}	1.0648	0.0708		1.0552	0.0704
α_{12}	2.1394	0.1000		2.1269	0.0993
α_{21}	1.3999	0.0782		1.3843	0.0775
α_{22}	1.8800	0.0921		1.8682	0.0915
α_{23}	2.7103	0.1293		2.7036	0.1286

From tables, it can be seen that the minimum chi-squared estimator, $\widehat{\theta}_{\phi_{(1)}}$ is the best. In fact, it is better than the classical maximum likelihood estimator, $\widehat{\theta}_{\phi_{(0)}}$ proposed by Dale(1986). By brevity we do not present the results for all the models considered by Dale but the conclusion is the same, the best choice is the minimum chi-squared estimator. Of course, it is necessary further study through a simulation design.

Acknowledgments: This work was partially supported by Grants MTM2009-06997 and BSCH–UCM (2008-910707).

References

Agresti, A. (2002). *Categorical Data Analysis.* Second Edition. Wiley, New York.

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics.* London : Chapman and Hall.

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

Dale, J.R. (1986). Global Cross-Ratio Models for Bivariate, Discrete, Ordered Responses. *Biometrics*, Vol. 42, N^o 4, pp 909-917.

Diggle, P.J., Heagerty, P.J., Liang, K.Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data.* Second Edition. Oxford University Press, Oxford.

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

Pardo, L. (2006). *Statistical Inference Based on Divergence Measures.* Chapman & Hall.

Verbeke,G. and Molenberghs,G. (1997). *Linear Mixed Models in Practice: A SAS-Oriented Approach.* Lecture Notes in Statistics 126. New York: Springer Verlag.

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

On composite likelihood estimation of a multivariate Poisson INAR(1) model

Xanthi Pedeli[1], Dimitris Karlis[1]

[1] Department of Statistics, Athens University of Economics and Business

Abstract: Multivariate count time series data occur in several different disciplines like epidemiology, criminology, engineering and marketing, just to name a few. However, relative literature is rather limited. The existing models have several limitations and they do not lead to models with well specified marginals. Hence inference can be difficult with standard methods like maximum likelihood and some alternatives, like composite likelihood, should be considered.
We define a multivariate INAR(1) model (MINAR(1)) and provide basic statistical properties of different estimators. Our main focus is placed on composite likelihood estimation, a recently popular and elegant method which makes the estimation problem computationally less demanding at the cost of some loss of efficiency.

Keywords: MINAR model; count data; multivariate Poisson distribution; composite likelihood.

1 Introduction

Consider the case where multivariate count data are observed in several successive time points. For example, consider a basket with different products where the data refer to the number of purchases for each product in a daily basis for one year. Or the number of admissions in a hospital for three different age groups observed in a large time period. In such cases one needs to specify a time series model appropriate for multivariate count data. In this paper we define such a model and provide estimation methods based on the idea of composite likelihood.

2 The Multivariate INAR(1) Process

Let $\mathbf{X}$ and $\mathbf{R}$ be non-negative integer-valued random n-vectors and let $\mathbf{A}$ be a $n \times n$ diagonal matrix with independent elements $\{\alpha_{jj}\}_{j=1}^{n}$. The multivariate integer-valued autoregressive process of order 1 (MINAR(1))

can be defined as

$$\mathbf{X}_t = \mathbf{A}\circ\mathbf{X}_{t-1}+\mathbf{R}_t = \begin{bmatrix} \alpha_1 & 0 & \cdots & 0 \\ 0 & \alpha_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha_n \end{bmatrix} \circ \begin{bmatrix} X_{1,t-1} \\ X_{2,t-1} \\ \vdots \\ X_{n,t-1} \end{bmatrix} + \begin{bmatrix} R_{1t} \\ R_{2t} \\ \vdots \\ R_{nt} \end{bmatrix}, \ t \in Z \tag{1}$$

where $\mathbf{A}\circ$ is a matricial operation which acts as the usual matrix multiplication keeping in the same time the properties of the binomial thinning operation. Obviously, the jth element of the bivariate series, $j = 1, \ldots, n$, is given by

$$X_{jt} = \alpha_j \circ X_{j,t-1} + R_{jt} \tag{2}$$

where $\alpha \circ X = \sum_{i=1}^{X} Y_i = Y$, with Y_i being a sequence of *iid* Bernoulli random variables such that $P(Y_i = 1) = \alpha = 1 - P(Y_i = 0)$ and $\alpha \in [0, 1]$ (Steutel and van Harn, 1979). The $\alpha \circ X$ operator represents binomial thinning of X such that each of the X individuals either "survives"with probability α or "dies"with probability $1 - \alpha$. Similarly to the INAR(1) process, the value of the MINAR(1) process at time t, denoted by $\mathbf{X}_t$, consists of two parts. The first part is comprised by the survivors of the elements of the process at the preceding point in time $t - 1$, denoted by $\mathbf{X}_{t-1}$, each with probability of survival equal to $\mathbf{A}$. The elements $\mathbf{R}_t$ which entered the system in the interval $(t-1, t]$ are usually called as innovations. Assuming independence between and within the thinning operations and $\{R_{jt}\}$ an *iid* sequence with mean λ_j and variance $\sigma_j^2 = \upsilon_j\lambda_j$, $\upsilon_j > 0, j = 1, \ldots, n$, the unconditional first and second order moments of the MINAR(1) process are proved to be similar to those of the BINAR(1) model (see Pedeli and Karlis, 2010). Moreover, in accordance to the BINAR(1) process, dependence between any two series that comprise the MINAR(1) process is introduced by allowing for dependence between the respective innovation terms. Thus, if $cov(R_{it}, R_{jt}) = \lambda_{ij}$, it can be shown that

$$cov(X_{i,t+h}, X_{jt}) = \frac{\alpha_i^h}{1 - \alpha_i\alpha_j}\lambda_{ij} \ ; \quad h = 1, 2, \ldots, \quad i \neq j \tag{3}$$

3 The Multivariate Poisson INAR(1) Model

We start with multivariate Poisson distribution used in Karlis and Meligkotsidou (2005). The model allows for a different covariance term for each pair of variables and thus it can be considered as a discrete counterpart of the multivariate normal distribution, suitable for multivariate count data. In this section, we use the idea of the multivariate Poisson model with two-way covariance structure in order to define a paricular joint mass function for the innovations of the MINAR(1) process.

More specifically, using the specification of Jost et al. (2006) for the vector of innovations $\mathbf{R} = (R_1 \dots R_n)^T$ we generate a multivariate Poisson INAR(1) model. Denoting by λ_j, $j = 1, \dots, n$, the j-th element of the mean vector $E(\mathbf{R})$ (or equivalently the j,j-th element of the variance-covariance matrix $Var(\mathbf{R})$) and by λ_{ij}, $i,j = 1, \dots, n$, $i \neq j$, the i,j-th element of the variance-covariance matrix $Var(\mathbf{R})$, we obtain the first- and second-order moments of the multivariate Poisson INAR(1) model. Namely, the vector of expectations $\boldsymbol{\mu}_{\mathbf{X}_t} = E(\mathbf{X}_t)$ with elements $\mu_{X_{jt}} = \lambda_j/(1-\alpha_j)$, $j = 1, \dots, n$; the variance-covariance matrix $\boldsymbol{\gamma}_{\mathbf{X}_t}(h)$ with diagonal elements $cov(X_{j,t+h}, X_{jt}) = \alpha_j^h \lambda_j/(1-\alpha_j)$ and off-diagonal elements $cov(X_{i,t+h}, X_{jt}) = \alpha_i^h \lambda_{ij}/(1-\alpha_i\alpha_j)$, $i \neq j$, $h = 0, 1, \dots$; and the correlation matrix $\boldsymbol{\rho}_{\mathbf{X}_t}(h)$ with diagonal elements $corr(X_{j,t+h}, X_{jt}) = \alpha_j^h$ and off-diagonal elements $corr(X_{i,t+h}, X_{jt}) = \alpha_i^h \lambda_{ij} \sqrt{(1-\alpha_i)(1-\alpha_j)}/\sqrt{\lambda_i \lambda_j}(1-\alpha_i\alpha_j)$, $i \neq j$, $h = 0, 1, \dots$

4 Estimation based on composite likelihood method

For conditional maximum likelihood estimation one needs to evaluate complicated multivariate probability mass functions. The computational complexity of the maximum likelihood approach augments with dimensional increase, rendering parameters estimation a laborious task. In order to overcome such practical difficulties, the concept of composite likelihood estimation may be used. Composite likelihood methods are based on the idea of constructing lower dimensional score functions that still contain enough information about the structure considered but they are computationally more tractable (see, e.g. Varin, 2008).
Following Jost et al. (2006), we define the bivariate marginal log-likelihood function between two random elements X_u and X_v as

$$\ell_{uv}(\boldsymbol{\theta}) = \frac{1}{T} \sum_{t=1}^{T} log f_{X_u, X_v}(x_{ut}, x_{vt}|\boldsymbol{\theta}) \tag{4}$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\lambda})^T$ is the vector of unknown parameters and

$$\begin{aligned} f_{X_u,X_v}(x_{ut}, x_{vt}|\boldsymbol{\theta}) &= \sum_{k_u=0}^{s_1} \sum_{k_v=0}^{s_2} \binom{x_{u,t-1}}{x_{ut}-k_u} \alpha_u^{x_{ut}-k_u} (1-\alpha_u)^{x_{u,t-1}-x_{ut}+k_u} \\ &\times \binom{x_{v,t-1}}{x_{vt}-k_v} \alpha_v^{x_{vt}-k_v} (1-\alpha_v)^{x_{v,t-1}-x_{vt}+k_v} \\ &\times exp(\lambda_u + \lambda_v - \lambda_{uv}) \\ &\times \sum_{m_{uv}=0}^{min(k_u,k_v)} \frac{(\lambda_u - \lambda_{uv})^{k_u-m_{uv}}}{(k_u-m_{uv})!} \frac{(\lambda_v - \lambda_{uv})^{k_v-m_{uv}}}{(k_v-m_{uv})!} \frac{\lambda_{uv}^{m_{uv}}}{m_{uv}!} \end{aligned} \tag{5}$$

where $s_1 = min(x_{ut}, x_{u,t-1})$, $s_2 = min(x_{vt}, x_{v,t-1})$.

The composite log-likelihood function $\ell(\boldsymbol{\theta})$ is further defined as the sum of all bivariate log-likelihood functions, i.e.

$$\ell(\boldsymbol{\theta}) = \sum_{u=1}^{n-1} \sum_{\upsilon=u+1}^{n} w_{u\upsilon} \ell_{u\upsilon}(\boldsymbol{\theta}) \tag{6}$$

where $w_{u\upsilon}$ is a constant weight for $\ell_{u\upsilon}(\boldsymbol{\theta})$. For simplicity, it is common to set $w_{u\upsilon} = 1$ for $1 \leq u \leq \upsilon \leq m$.

One can see that the specification of the bivariate probability mass function (4) is a straightforward generalization of the conditional density of the Poisson BINAR(1) model. Moreover, (4) is a convolution of two binomials and a bivariate Poisson distribution with appropriately defined parameters. This result is also in accordance with the specification of the conditional density for the Poisson BINAR(1) model.

Finally, the computational benefit of the pairwise composite likelihood approach over the method of maximum likelihood, lies on the significant diminution of the number of required summations. In particular, the maximum likelihood estimator uses a multivariate probability mass function which involves $\frac{1}{2}n(n-1)$ nested summations, while for the calculation of the composite log-likelihood only three summations are required.

We have run simulation experiments to examine the properties of this approach and the loss of efficiency with respect to the full likelihood approach. Also real data application will be used to illustrate the application potential of the method.

References

Jost, T.A., Brcich, R.F., and Zoubir, A.M. (2006). Estimating the parameters of the multivariate Poisson distribution using the composite likelihood concept. In: *In Proceedings of the 31st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Toulouse, France. Vol 3, pp. 656-659.

Karlis, D., and Meligkotsidou, L. (2005). Multivariate Poisson regression with covariance structure. *Statistics and Computing*, **15**, 255-265.

Pedeli, X., and Karlis, D. (2010). A bivariate INAR(1) process with application. *Statistical Modelling: An International Journal*, (to appear).

Steutel, F.W., and van Harn, K. (1979). Discrete Analogues of Self-Decomposability and Stability. *The Annals of Probability*, **7(5)**, 893-899.

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis*, 92, 1-28.

Approximate Bayesian Computation using Auxiliary Model Based Estimates

Anthony N Pettitt[1], Christopher C Drovandi[1], Malcolm J Faddy[1]

[1] Mathematical Sciences, Queensland University of Technology. 2 George St, Brisbane, QLD, Australia, 4001. Email: a.pettitt@qut.edu.au.

Abstract: We present a novel approach for developing summary statistics for use in approximate Bayesian computation (ABC) algorithms using indirect inference. We embed this approach within a sequential Monte Carlo algorithm that is completely adaptive. This methodological development was motivated by an application involving data on macroparasite population evolution modelled with a trivariate Markov process. The main objective of the analysis is to compare inferences on the Markov process when considering two different indirect models. The two indirect models are based on a Beta-Binomial model and a three component mixture of Binomials, with the former providing a better fit to the observed data.

Keywords: Approximate Bayesian computation, Beta-Binomial model, Binomial mixture model, Indirect inference, Markov process, Sequential Monte Carlo.

1 Introduction

In approximate Bayesian computation (ABC), we seek to make inferences about the parameters of the posterior distribution when the likelihood function is computationally intractable. While the likelihood function itself cannot be computed easily, it is assumed that simulation from the model is relatively straightforward. The likelihood is replaced by a comparison of p summary statistics, $S(\cdot) = [S_1(\cdot), \ldots, S_p(\cdot)]^T$, of the observed and simulated data using a distance function, $\rho(y, x)$

$$\rho(y, x) \quad = \quad \|S(y) - S(x)\|.$$

ABC is particularly effective when the statistics are sufficient. However, in many applications sufficient statistics are not available and the practitioner must resort to a selection of carefully chosen data summaries.
In this paper we investigate an alternative approach to obtaining summary statistics based on indirect inference (Heggland and Frigessi 2004). In indirect inference an auxiliary model is proposed whose likelihood function is tractable and provides a good description of the data. The objective is to

search for parameter values of the model of interest that produce simulated data that lead to auxiliary parameters close to those based on maximum likelihood of the original data. Therefore a comparison of summary statistics involves computing a distance between such auxiliary parameters.
We consider a stochastic process model developed by Riley et al (2003) for a macroparasite population within a host. A Beta-Binomial model or a Binomial mixture is employed as an auxiliary model to provide a description of the data, while the stochastic model encapsulates the biological system which drives the observed data. We investigate the sensitivity of the inferences on the Markov process model to the indirect model. In particular we analyse any inefficiencies by introducing a three component Binomial mixture, which does not fit the data as well as the Beta-Binomial model.

2 Data and Modelling

Here the data is described as well as the stochastic process model of Riley et al (2003) used to explain the data. We also outline the auxiliary models.

2.1 Data

The data consist of mature parasite counts at particular autopsy times for 212 hosts (Denham et al 1972). Each host was injected with roughly 100 or 200 larvae and necropsy time ranged between 24 and 1193 days after the initial infection. The data are in the form of proportions (the mature count divided by the initial infection). From Figure 1 there is clear evidence of overdispersion, which a Binomial distribution alone cannot describe.

2.2 Markov Process Model

The following stochastic model was developed by Riley et al (2003) to help explain the population dynamics of *Brugia pahangi*. At time t any host is described by three random variables $\{M(t),\ L(t),\ I(t)\}$, where $M(t)$ is the number of mature parasites, $L(t)$ is the number of larvae and $I(t)$ is a discrete immunity variable. Initially cats are infected with L_I larvae and after a certain time the hosts are autopsied and the number of mature parasites are recorded. It is assumed that larvae can mature at a rate of γ per larva per day. Larvae die at a rate $\mu_L + \beta I(t)$ per larva where μ_L represents natural death of larvae and β describes the death of larvae due to the immune response of the host. The acquisition of immunity occurs at rate $\nu L(t)$, and a host loses immunity at a rate μ_I per unit of immunity. Mature parasites die at a rate of μ_M adults per day. In its deterministic form, the above model can be re-written as a set of differential equations

$$\frac{dL}{dt} = -\mu_L L - \beta IL - \gamma L,\ \frac{dM}{dt} = \gamma L - \mu_M M,\ \frac{dI}{dt} = \nu L - \mu_I I.$$

We consider the stochastic version of this model via a continuous time discrete trivariate Markov process as developed by Riley et al (2003). Data can be simulated from the model using the algorithm of Gillespie (1977). We consider $\mu_M = 0.0015$ and $\gamma = 0.04$ fixed as per Riley et al (2003).

2.3 Auxiliary Models

For the auxiliary model we propose a Beta-Binomial model, which contains an extra parameter to capture the dispersion. More specifically, the i^{th} observation has a Beta-Binomial distribution with Beta parameters, α_i and β_i. It is convenient to use a reparameterisation in terms of the proportion, $p_i = \alpha_i/(\alpha_i + \beta_i)$, and overdispersion, $\theta_i = 1/(\alpha_i + \beta_i)$, so that the mean and variance are given by $l_i p_i$ and $l_i p_i(1-p_i)(1+(\theta_i/(1+\theta_i))(l_i-1))$. We relate these parameters to the necropsy time, t_i, and initial larvae burden, l_i, through the following functions chosen to optimise the fit to the data

$$\begin{aligned} \text{logit}(p_i) &= \beta_0 + \beta_1 \log(t_i) + \beta_2 \log(t_i)^2 \\ \log(\theta_i) &= \begin{cases} \eta_{100}, \text{ if } l_i \approx 100 \\ \eta_{200}, \text{ if } l_i \approx 200 \end{cases} . \end{aligned}$$

We also consider an alternative auxiliary model based on a three component Binomial mixture. This model was chosen purposefully as it still provides a reasonable description of the data but does not fit the data as well as the Beta Binomial model. The ith observation has the density

$$f(m_i|\Theta) = \binom{l_i}{m_i} \sum_{k=1}^{3} w_k (\theta_i^k)^{m_i} (1-\theta_i^k)^{l_i - m_i},$$

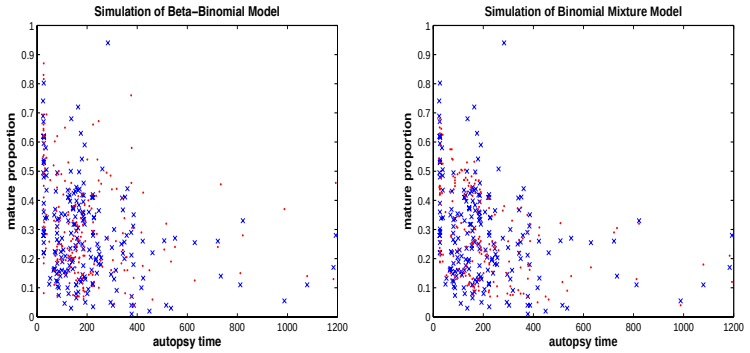
where $w_3 = 1 - w_1 - w_2$. We reparameterise the θ_i^k, $\text{logit}(\theta_i^k) = \gamma_0^k + \gamma_1 \log(t_i)$, so that each component has the same slope but a different intercept. Therefore this model has six parameters, $\Theta = (w_1, w_2, \gamma_0^1, \gamma_0^2, \gamma_0^3, \gamma_1)$. The Beta-Binomial model provides a more optimal fit, with an improvement of about 170 points in the loglikelihood using one less parameter. From Figure 1 it is clear that the Beta-Binomial is explaining more variability. Furthermore, the Beta-Binomial simulations are spread across the range of observed matures while the mixture simulations are 'clumpy'.
Our main investigation is to determine whether using the mixture auxiliary model leads to inefficient parameter estimates of the stochastic process model compared to when using the Beta-Binomial auxiliary model. Or, are inferences sensitive to the choice of the auxiliary model and how much effort should be spent on finding a well-fitting indirect model?

3 ABC using Indirect Inference

We consider a sequential Monte Carlo ABC (Sisson et al 2007) algorithm to sample from the sequence of targets

$$\pi(\theta, x|\rho(y,x) \leq \epsilon_t) \propto f(x|\theta)\pi(\theta) 1_{\rho(y,x)\leq\epsilon_t} \text{ for } t = 1, \ldots, T.$$

FIGURE 1. A typical simulation from the Beta-Binomial (left) and the Binomial mixture (right) models. A cross denotes observed and a dot denotes simulated.



Our ABC algorithm is based upon the SMC ABC replenishment algorithm of Drovandi and Pettitt (2010). Here N particles are traversed through the sequence of target distributions. The algorithm determines the sequence of tolerances adaptively by dropping a proportion, α, of the particles with the highest discrepancy value. The population is replenished by resampling from the remaining particles. Diversity is ensured by moving the particles according to an MCMC kernel invariant for the current target. The proposal distribution of this MCMC step is also updated dynamically. We iterate the MCMC kernel sufficiently to ensure that each particle gets moved with a theoretical probability of $1 - c$ (with c set small). Our algorithm differs from Sisson et al (2009) and Beaumont et al (2009) since they use a forward kernel and also require pre-specification of the sequence of tolerances.
ABC with indirect inference requires an extra step. After data are simulated from the model, an auxiliary model is fitted to the data. The parameter estimates of this auxiliary model become the simulated summary statistics, θ_a^x, which are then compared to the observed summary statistics, $\hat{\theta}_a$.

4 Results

We inferred the parameters of the stochastic model successfully using the indirect inference approach with both auxiliary models. In the ABC algorithm we used $N = 1000$ particles, dropped half the particles with the worst discrepancy, $\alpha = 0.5$, and iterated the MCMC kernel so that theoretically 99% of the particles are moved, $c = 0.01$. For both cases the process was stopped when the MCMC kernel had about a 3% acceptance rate.
Parameter summaries of the Markov process model when applying each of the auxiliary models is presented in Table 1. Unfortunately the parameters μ_I and β are imprecisely estimated. This occurred since only mature counts are available and the immunity variable in simulations mostly takes a value

TABLE 1. Posterior summaries. Shown are the posterior mode, mean, standard deviation and the (2.5%,50%,97.5%) quantiles. BB = Beta-Binomial indirect model and BM = Binomial mixture indirect model. † estimates for these parameters have been multiplied by 100.

model	param	mode	mean	std dev	(2.5%,50%,97.5%)
BB	ν†	0.13	0.13	0.03	(0.07,0.13,0.20)
BB	μ_I	1.08	1.03	0.47	(0.15,1.02,1.88)
BB	μ_L†	0.55	0.85	0.60	(0.04,0.73,2.35)
BB	β	1.34	1.20	0.44	(0.34,1.22,1.96)
BM	ν†	0.08	0.11	0.04	(0.05,0.11,0.22)
BM	μ_I	1.05	1.03	0.46	(0.20,1.03,1.89)
BM	μ_L†	2.44	2.07	0.68	(0.37,2.22,3.05)
BM	β	1.03	1.18	0.43	(0.40,1.17,1.95)

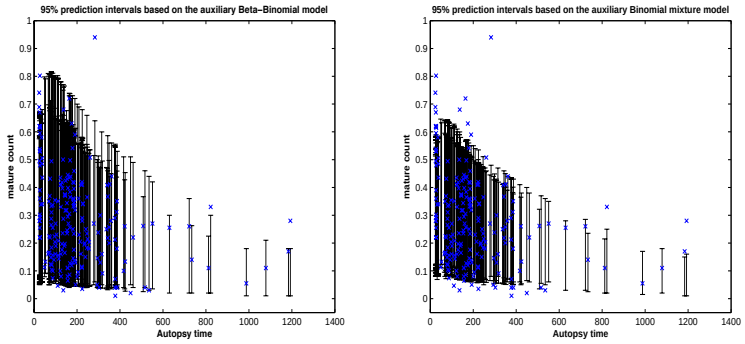
no higher than 1 and is short lived. This meant that the parameters were sensitive to the prior but the ratio μ_I/β was relatively less sensitive.
The parameters ν and μ_L were precisely estimated. It can be seen from the Table that the posterior summaries for ν were similar regardless of which auxiliary model was applied, with a smaller variance when the Beta-Binomial model was used. The posterior summaries for μ_L were more dependent on the indirect model. The posterior for μ_L when using the Beta-Binomial distribution is shifted, tighter and is skew to the right compared with the posterior when the Binomial mixture is used.
To compare the results from the two auxiliary models we produced predictions of the Markov process model based on the posterior modes in Table 1. We approximated 95% prediction intervals based on each auxiliary model, shown in Figure 2. It is clear that predictions from the stochastic model using Beta-Binomial auxiliary estimates account for more variability in the data. However, it appears that the 'clumpiness' of the Binomial mixture fit does not cause any problems as the stochastic model cannot predict such an effect. However, the most important summary would seem to be the range of the data at each time point, which the Beta-Binomial model explains better than the Binomial mixture model.

Acknowledgments: The authors would like to thank Edwin Michael and David Denham for access to the data. The authors also wish to thank Chris Glasbey and Steven Riley.

References

Beaumont, M. A., Cornuet, J-M, Marin, J-M and Robert, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika*, **96**, 983-990.

FIGURE 2. 95% predictions intervals based on the posterior modes of the stochastic model when applying the Beta-Binomial (left) and the Binomial mixture (right) as auxiliary models.



Denham, D. A., Ponnudurai, T., Nelson, G. S., Guy, F. and Rogers, R. (1972). Studies with *Brugia pahangi*. I. Parasitological observations on primary infections of cats (Felis catus). macroparasite models. *International Journal for Parasitology*, **2**, 239-247.

Drovandi, C. C. and Pettitt, A. N. (2010). Estimation of Parameters for Macroparasite Population Evolution using Approximate Bayesian Computation. *To appear in Biometrics*.

Gillespie, D. T. (1977). Exact Stochastic Simulation of Coupled Chemical Reactions. *The Journal of Physical Chemistry*, **81**, 2340-2361.

Heggland, K. and Frigessi, A. (2004). Estimating functions in indirect inference. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, **66**, 447-462.

Riley, S., Donnelly, C.A. and Ferguson, N.M. (2003). Robust parameter estimation techniques for stochastic within-host macroparasite models. *Journal of theoretical biology*, **225**, 419-430.

Sisson, S. A., Fan, Y. and Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 1760-1765.

Sisson, S. A., Fan, Y. and Tanaka, M. M. (2009). Correction for Sisson et al., sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America* , doi:10.1073/pnas.0908847106.

On probabilities of avalanches triggered by alpine skiers. An empirically driven decision strategy for backcountry skiers based on these probabilities.

Christian Pfeifer[1]

[1] Institut fr Statistik, Universitt Innsbruck, Universittsstrae 15, A–6020 Innsbruck

Abstract: This paper gives a decision strategy for backcountry skiers based on empirical probabilities. They are the result of a logistic regression model based on data of avalanche events and forecasts in Tyrol within three seasons (1999-2002).

Keywords: avalanche danger; decision strategy.

1 Introduction

In Austria, most fatal snow avalanche accidents are caused by skiers or snowboarders. 79 avalanche accidents (17 fatalities) were reported during the winter of 2001/02. 16 out of 17 these fatalities were caused by alpine skiers or snowboarders. By far the highest number of accidents took place in Tyrol (2001/02: 47 accidents/ 12 fatalities). However, it is rather difficult to predict the risk (=probability) of avalanche events on a backcountry ski slope under given conditions. About 10 years ago, the mountain guide Werner Munter suggested a quantitative method in order to estimate the risk of avalanche events. Assuming that the variables

- danger levels from the local avalanche information service (low=1 to very high=5),
- incline of the slope (three classes from flat to steep),
- aspect of the slope (north, south) and
- skiers behaviour

have an influence on the risk, he calculated a quantity which he calls "remaining risk". On the base of this quantity, he developed a strategy for backcountry skiers whether to go or not to go on a skiing tour (stop if "remaining risk" is larger than 1, see [1]). But Munter's quantity cannot be understood as a probability of avalanche events. Moreover, there is no

empirical evidence for his method because he does not take skiing incidents without avalanche accidents into account ([2]). At least, it is necessary to include some information on frequencies of skiers on slopes under specific conditions.

2 Statistical models

In Rothart and Pfeifer (2003), we present an approach which seems to be the first one where results on probabilities of avalanches triggered by skiers have been given. This consists in modeling the counts y_i of avalanche events in each class of incline and aspect for days i with avalanche reports from the Tyrolean avalanche information service (Lawinenwarndienst Tirol).

$$\log(y_i) = \texttt{LWS} + \texttt{NEIG} + \texttt{EXPOS} + \texttt{WOENDE} + \texttt{TOURV}$$

Beside danger level `LWS`, incline of slope `NEIG` and aspect of slope `EXPOS`, we took the qualitative variables skiing conditions `TOURV` and day of the week `WOENDE` into consideration. There is some evidence that frequencies of skiers on slope strongly depend on weather and snow conditions and on the days of the week (weekend, working days). We used accident data and avalanche forecasts in Tyrol within the seasons 1999-2002 reported by the Tyrolean avalanche information service (497 days of observation). However, because avalanche accidents are expected to be rather rare this simple Poisson model shows strong underdispersion. To overcome this misspecification we proposed the following statistical models (see Pfeifer and Rothart (2004)):

- Zero inflated Poisson models (**ZIP**): The observations y_i are expected to be drawn from a mixture of a Bernoulli and Poisson distribution.
- Zero altered Poisson models (**ZAP**): The observations y_i are expected to come from mixture that is zero with probability one in the first component and a truncated Poisson in the second component.

Using this models for counts with extra zeros seems to increase the goodness of fit of the Poisson model. The predicted probabilities are slightly lower than in the Poisson case. However, for the purpose of getting predicted probabilities there is no essential difference between specially built Poisson models and the logistic model in the following.

2.1 Logistic Model

In this approach, we propose a logistic regression model for reasons of simplification (no vs. one or more accidents as dependent variable for days i with avalanche reports from the Tyrolean avalanche information service),

in order to estimate the probabilities **p** in question. As mentioned before, the variables `WOENDE` and `TOURV` are taken into the model in order to control effects due to different backcountry skier frequencies.

$$\text{logit}(\mathbf{p}) = \texttt{LWS} + \texttt{NEIG} + \texttt{EXPOS} + \texttt{WOENDE} + \texttt{TOURV}$$

Table 1 gives the estimated parameters of this logistic model.

	Value	Std. Error	t value
(Intercept)	-7.2283515	0.6393772	-11.3053003
LWS	0.9116466	0.1781101	5.1184440
NEIG	0.8324917	0.1464717	5.6836361
as.factor(EXPOS)	-0.5777892	0.2153850	-2.6825877
as.factor(WOENDE)	0.4006435	0.2147187	1.8658994
as.factor(TOURV)2	-0.1226823	0.2903134	-0.4225857
as.factor(TOURV)3	-0.9364877	0.3808698	-2.4588132

Table 1: Estimated parameters of the logistic regression model

Figure 1 shows the distribution of the predictions (=probabilities) of the logistic model fitted to the avalanche data.

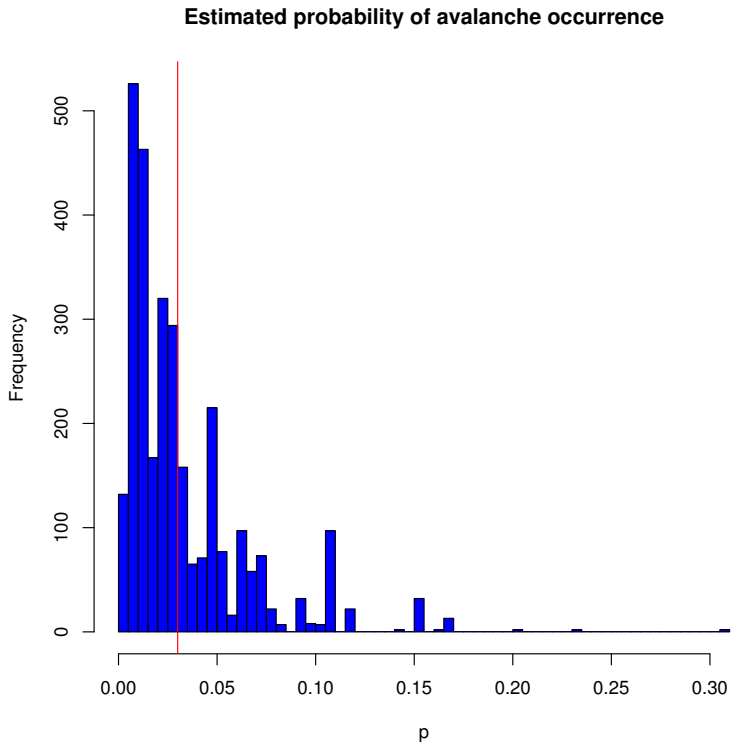


Figure 1: Histogram of estimated probabilities and cut-point for decision stop/go

3 Decision strategy

Further on, we try to establish a decision strategy for backcountry skiers based on empirical/statistical arguments. We have to fix a limit probability p^* which represents the cut-point for the decision whether to stop or to go. If the probability triggering an avalanche is smaller (higher) than p^*, then the backcountry skier would decide to go on the tour (to stop the tour). We choose this cut-point in such a way: For given p, calculate the 2×2 contingency table based on the variables avalanche occurrence yes/no and decision stop/go and quantify the dependence measure $\chi^2(p)$. Identify $p*$ in such a way that maximizes the function $\chi^2(p)$, $0 \leq p \leq 1$. As a result of this we get the cut-point p^* where the dependence between the variables avalanche occurrence and decision stop/go is a maximum. The vertical line in Fig. 1 indicates this point p^* (equal to 0.03) in our case.
Finally we are able to suggest the decision strategy depending on the variables danger level, incline of the slope and aspect of the slope in Figs. 2 and 3. The rows represent the three classes of the slope incline and the columns represent five classes of danger level. The meaning of the colours of the boxes is the following:

- green (go): relative frequency of predicted cases where to stop is equal to zero;
- yellow (attention): relative frequency of predicted cases where to stop is smaller than 50%;
- red (stop): relative frequency of predicted cases where to stop is larger than 50%.

4 Conclusion

For the purpose of obtaining acceptance from the avalanche research community, we tried to give an empirically driven decision strategy for backcountry skiers in this paper using rather simple statistical techniques. Our proposal is more or less comparable to Munter's method. But the aspect of the slope seems to be not that important as widely believed (if we notice the higher predicted cases in the yellow boxes of Fig. 3). Finally we recommend to to additional research (count data based on a random sample instead of qualitative data) in order to get more precise information on the frequencies of backcountry skiers on slopes.

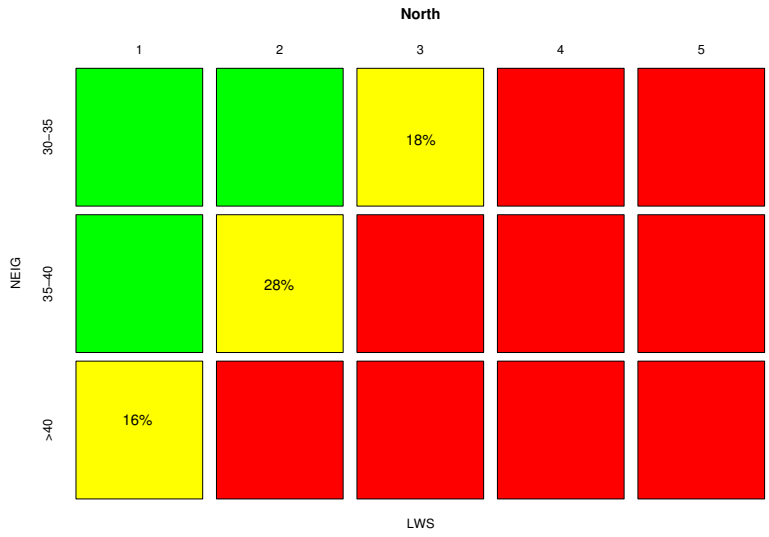


Figure 2: Decision strategy in the northern sector dependent on the danger level and the incline of the slope (go/green, yellow/attention and stop/red)

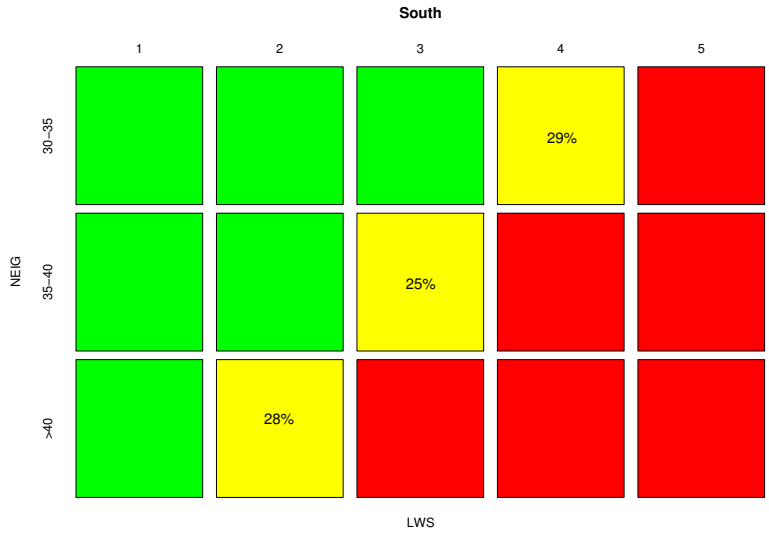


Figure 3: Decision strategy in the southern sector dependent on the danger level and the incline of the slope (go/green, yellow/attention and stop/red)

References

[1] Munter W. (1997): 3x3 Lawinen, Pohl & Schellhammer, Garmisch-Partenkirchen.

[2] Pfeifer C. Rothart V. (2002): Die Reduktionsmethode zur Beurteilung der Lawinengefahr fr Schitourengeher aus statistischer Sicht; Jahrbuch der Kuratoriums fr alpine Sicherheit 2002, Innsbruck.

[3] Pfeifer C. Rothart V. (2004): On probabilities of avalanches triggered by alpine skiers. An application of models for counts with extra zeros; Proceedings International Workshop of Statistical Modelling 2004 Florence

[4] Pfeifer C. (2009): On probabilities of avalanches triggered by alpine skiers. An empirically driven decision strategy for backcountry skiers based on these probabilities; Natural Hazards 2009; 48/3: 425-438

[5] Pfeifer C. (2009): On probabilities of avalanches triggered by alpine skiers. An empirically driven decision strategy for backcountry skiers based on these probabilities; presentation at International Snow Science Workshop ISSW2009 27.9.2009 - 2.10.2009 Davos Switzerland

[6] Rothart V. Pfeifer C. (2003): Neuere Methoden zur Beurteilung der Lawinengefahr fr Schitourengeher aus statistischer Sicht. Ein erstes statistisches Modell mit Informationen von Begehungsfrequenzen; presentation at sterreichische Statistiktage 2003 31.10.2003 Vienna

Address for correspondence

Christian Pfeifer
Institut fr Statistik, Universitt Innsbruck
Universittsstrae 15, A–6020 Innsbruck
E-Mail: christian.pfeifer@uibk.ac.at

Sensitivity analysis for incomplete continuous data

Frederico Z. Poleto[1], Geert Molenberghs[2], Carlos Daniel Paulino[3], Julio M. Singer[1]

[1] IME, Universidade de São Paulo, Caixa Postal 66281, São Paulo, SP, 05314-970, Brazil. E-mails: fpoleto@ime.usp.br and jmsinger@ime.usp.br.
[2] I-BioStat, Universiteit Hasselt, B-3590 Diepenbeek and Katholieke Universiteit Leuven, B-3000 Leuven, Belgium. E-mail: geert.molenberghs@uhasselt.be.
[3] IST, Universidade Técnica de Lisboa (and CEAUL-FCUL), Av. Rovisco Pais, Lisboa, 1049-001, Portugal. E-mail: dpaulino@math.ist.utl.pt.

Abstract: In studies with missing data, untestable assumptions are required to identify statistical models. Such assumptions are usually questionable and statisticians commonly bypass the problem via sensitivity analyses. Specifically for continuous data, previous work has been developed under assumptions of normality and/or employing hard-to-interpret sensitivity parameters. We derive a simple approach for estimating means, standard deviations and correlations avoiding any parametric distribution assumption for the outcomes. Adopting a pattern-mixture model parameterization, we employ non-identifiable means, standard deviations and correlations, or functions thereof, as sensitivity parameters, which, we believe, are more easily elicited by experts.

Keywords: Identifiability; Ignorance interval; Missing data; Pattern-mixture model; Uncertainty interval.

1 Motivating example

The U.S. News & World Report's Guide 1995 collected information from 1,302 American colleges. Here we focus on 3 variables: CSAT (average combined math and verbal Scholastic Assessment Test), GRADRAT (ratio between the number of graduating seniors and the number of enrolled students four years earlier $\times 100$) and an indicator of public vs. private colleges. The indicator was the only variable without missing values. We are interested in the following questions: i) do the public and private colleges have different mean CSAT? and ii) are CSAT and GRADRAT linearly correlated? Descriptive statistics related to the aforementioned objectives are displayed in Tables 1 and 2.

2 Univariate case

Let Y_i denote the measurement of the i-th unit of the study and R_i be an indicator variable taking on the value 1 if Y_i is observed and 0 otherwise, $i = 1, \ldots, n$. Using the pattern-mixture model parameterization and

TABLE 1. Counts, means and standard deviations (SD) for CSAT.

College administration	CSAT observed			CSAT missing		
	Count	Mean	SD	Count	Mean	SD
Public	251	945.3	107.5	219	?	?
Private	528	978.8	129.2	304	?	?

? denotes non-observed values

TABLE 2. Counts, means, standard deviations (SD) and correlations for GRADRAT and CSAT.

Missingness pattern			GRADRAT		CSAT		
GRADRAT	CSAT	Count	Mean	SD	Mean	SD	Correlation
Observed	Observed	731	62.0	18.5	974.0	123.0	0.594
Observed	Missing	472	57.7	19.0	?	?	?
Missing	Observed	48	?	?	876.6	93.5	?
Missing	Missing	51	?	?	?	?	?

? denotes non-observed values

conditional expectation properties, we get

$$\mu = \mathrm{E}(Y_i) = \gamma_1\mu_{(1)} + \gamma_0\mu_{(0)}, \quad (1)$$

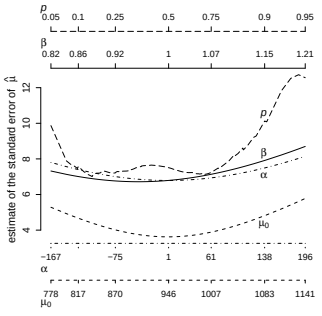
$$\sigma^2 = \mathrm{Var}(Y_i) = \gamma_1\sigma^2_{(1)} + \gamma_0\sigma^2_{(0)} + \gamma_1[\mu_{(1)} - \mu]^2 + \gamma_0[\mu_{(0)} - \mu]^2, \quad (2)$$

where $\gamma_r = P(R_i = r)$, $\mu_{(r)} = \mathrm{E}(Y_i|R_i = r)$ and $\sigma^2_{(r)} = \mathrm{Var}(Y_i|R_i = r)$, for $r = 0, 1$.

We can estimate γ_1 ($\gamma_0 = 1 - \gamma_1$), $\mu_{(1)}$ and $\sigma^2_{(1)}$ by their sample counterparts $\hat{\gamma}_1$ ($\hat{\gamma}_0$), $\hat{\mu}_{(1)}$ and $\hat{\sigma}^2_{(1)}$. Although $\mu_{(0)}$ and $\sigma^2_{(0)}$ are not identified from the observed data, if we set values for them we may obtain an unbiased estimate $\hat{\mu}(\mu_{(0)})$ of $\mu(\mu_{(0)})$ and a consistent estimate $\hat{\sigma}^2\big(\mu_{(0)}, \sigma^2_{(0)}\big)$ of $\sigma^2\big(\mu_{(0)}, \sigma^2_{(0)}\big)$. Therefore, $\omega = \mu_{(0)}$ or $\omega = \big(\mu_{(0)}, \sigma^2_{(0)}\big)$ are the so-called sensitivity parameters for the purpose of estimating μ or σ^2, respectively. The range of estimates obtained after repeating the analysis over a set $\boldsymbol{\Omega}$ of values for ω provides a Honestly Estimated Ignorance Region (HEIR). Likewise, the union of $100(1 - \alpha)\%$ confidence regions obtained for each ω provides a $100(1 - \alpha)\%$ Estimated Uncertainty Region (EURO). In the same way that standard errors and confidence regions quantify statistical imprecision due to sampling, ignorance regions measure the statistical ignorance on account of deficiencies of the observation process, like missing data, and the uncertainty region assesses the statistical uncertainty caused by the combination of imprecision and ignorance. Vansteelandt et al. (2006) consider a formal approach to the problem and provide appropriate definitions of consistency and coverage for these regions. They show how to construct EUROs for a scalar parameter π according to each definition of

the uncertainty region with uncertainty level $100(1-\alpha)\%$: (i) strong EUROs cover $\pi(\omega)$ simultaneously for all $\omega \in \mathbf{\Omega}$ with at least $100(1-\alpha)\%$ probability, (ii) pointwise EUROs cover $\pi(\omega)$ uniformly over $\omega \in \mathbf{\Omega}$ with at least $100(1-\alpha)\%$ probability, and (iii) weak EUROs have an expected overlap with the ignorance region of at least $100(1-\alpha)\%$.

For categorical missing data, the set $\mathbf{\Omega}$ may cover an in-depth grid of the whole parameter space of ω, but for continuous data, this strategy is clearly not feasible. Therefore, when there is no prior information to choose $\mathbf{\Omega}$ under a certain parameterization, the elicitation task may become easier under another. For example, in lieu of using $\mu_{(0)}$ as sensitivity parameter, we may prefer to use α, β or p, where $\mu_{(0)} = \alpha + \mu_{(1)}$, $\mu_{(0)} = \beta\mu_{(1)}$ and $\mu_{(0)} = F^{-1}_{(1)}(p)$ (the p-th quantile of the theoretical distribution of the observed units). The variance of the estimator of μ depends upon which sensitivity parameter strategy is used (expressions not shown) as portrayed in Figure 1, which contains estimates of standard errors of $\hat{\mu}$ for the data in Table 1. The 4 horizontal axes indicate equivalences among the 4 sensitivity parameters for estimating μ; for instance, $p = 0.9$, $\beta \cong 1.15$, $\alpha \cong 138$ and $\mu_0 \cong 1,083$ lead to the same $\hat{\mu}$ for public colleges, but different estimates of its standard error. Looking at the estimation of σ in (2), it may be more meaningful to work with λ in $\sigma^2_{(0)} = \lambda^2\sigma^2_{(1)}$ than with $\sigma_{(0)}$. A simple way to obtain an estimate of the variance of $\hat{\sigma}$, whether we use $\left(\mu_{(0)}, \sigma^2_{(0)}\right)$ or any other parameterization, is to employ the nonparametric bootstrap. In Table 3 we display estimated intervals for the mean and standard deviation of CSAT for public and private colleges and also for the corresponding mean CSAT difference using $\mathbf{\Omega} = [0.90; 1.00]$ for β and $\mathbf{\Omega} = [0.80; 1.25]$ for λ. The results do not allow us to conclude whether there is a difference between the means.



(a) (b)

FIGURE 1. Estimates of standard errors of $\hat{\mu}$ using $\mu_{(0)}$, α, β or p as sensitivity parameter for (a) public and (b) private colleges.

TABLE 3. HEIR and 95% weak, pointwise and strong EUROs for CSAT.

Admin.	Param.	HEIR	Weak	Pointwise	Strong
Pub.	μ	[901.2; 945.3]	[897.0; 949.5]	[890.0; 956.4]	[887.8; 958.6]
	σ	[98.0; 129.6]	[95.3; 132.8]	[90.7; 138.0]	[89.3; 139.7]
Priv.	μ	[943.0; 978.8]	[939.3; 982.4]	[933.7; 988.0]	[931.9; 989.8]
	σ	[120.4; 149.5]	[117.8; 152.6]	[113.5; 157.5]	[112.1; 159.0]
$\mu_{(\text{Priv.})} - \mu_{(\text{Pub.})}$		[−2.3; 77.5]	[−5.9; 81.2]	[−16.8; 92.1]	[−19.6; 94.9]

3 Multivariate case

Using the multivariate versions of (1) and (2), the analysis may be carried out along similar lines as in previous section. The challenges here are that there are additional options for the parameterization and that the number of sensitivity parameters may increase exponentially depending on the missingness patterns and number of variables. In the analysis of Table 4, for example, we conclude that GRADRAT and CSAT are positively linearly correlated; the magnitude of the correlation, however, is difficult to assess, given the ignorance caused by the missing data.

TABLE 4. HEIR and 95% weak, pointwise and strong EUROs.

	GRADRAT		CSAT		
Interval	Mean	SD	Mean	SD	Correlation
HEIR	[59.9; 60.4]	[18.6; 19.3]	[929.1; 968.0]	[114.3; 144.8]	[0.288; 0.795]
Weak	[59.1; 61.2]	[18.2; 19.7]	[927.4; 969.7]	[112.8; 146.6]	[0.300; 0.786]
Pointwise	[59.0; 61.3]	[18.0; 19.8]	[922.3; 975.1]	[108.8; 151.1]	[0.262; 0.817]
Strong	[58.9; 61.4]	[17.9; 20.0]	[921.0; 976.5]	[107.7; 152.3]	[0.257; 0.821]

Acknowledgments: We gratefully acknowledge the financial supports to this research: Frederico Z. Poleto and Julio M. Singer, from CAPES, CNPq and FAPESP, Brazil; Geert Molenberghs, from the IAP research Network P6/03 of the Belgian Government (Belgian Science Policy); Carlos Daniel Paulino, from FCT through the research centre CEAUL-FCUL, Portugal.

References

Vansteelandt, S., Goetghebeur, E., Kenward, M.G., and Molenberghs, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, **16**, 953-979.

Estimating biologically plausible relationships between air pollution and health

Helen Powell[1], Duncan Lee[1], Adrian Bowman[1]

[1] Department of Statistics, 15 University Gardens, University of Glasgow, Glasgow, G12 8QQ. Contact: *h.powell@stats.gla.ac.uk*

Abstract: This paper describes the construction of biologically plausible dose-response curves, in studies that estimate the short-term effects of air pollution on human health. Our methods are applied to a study in Glasgow, Scotland, between 2000 and 2007.

Keywords: Air pollution; monotonic dose-response relationship; respiratory health.

1 Background

The effects of air pollution concentrations on human health can be estimated using ecological time-series studies (see for example Dominici *et al* (2002)), which comprise daily data for the population living within an extended urban area. The response data, $\mathbf{y} = (y_1, \ldots, y_n)_{n\times 1}$, are daily counts of mortality or morbidity outcomes, which are related to average air pollution concentrations, $\mathbf{z} = (z_1, \ldots, z_n)_{n\times 1}$, and other covariates, $X = (\mathbf{x}_1^{\mathrm{T}}, \ldots, \mathbf{x}_n^{\mathrm{T}})_{n\times p}$. A general model for these data is given by

$$\begin{aligned} Y_t &\sim \text{Poisson}(\mu_t) \quad \text{for } t = 1, \ldots, n, \\ \ln(\mu_t) &= \mathbf{x}_t^{\mathrm{T}}\alpha + g(z_{t-k}), \end{aligned} \tag{1}$$

where the pollutant has been lagged by k days, and g represents the shape of its relationship to health. The majority of studies estimate a linear relationship between air pollution and health (e.g. $g(z_{t-k}) = z_{t-k}\beta$) for simplicity, although a number (e.g. Dominici *et al* (2002)) have estimated non-linear dose-response curves. However, these dose-response curves are typically unconstrained and estimated using smoothing or penalised splines, meaning that non-biologically plausible curves can occur. For example, for some levels of pollution the estimated health effects may decrease for increasing concentrations. Therefore this paper proposes a method for estimating biologically plausible dose-response curves for air pollution and health studies.

2 Methods

To construct a biologically plausible dose-response curve Shaddick *et al* (2008) suggest that it must satisfy the following properties: (i) increasing monotonicity; (ii) smoothness (thrice differentiability); and (iii) $g(0) = 0$, which together enforce the dose-response curve to be non-negative (i.e. pollution exposure cannot be beneficial to health). They also suggest that the curve should be bounded from above, but as relatively low levels of pollution are observed in the majority of cities worldwide, an upper limit on the health effects may not be observable from the available data. We therefore do not consider this constraint. Our proposed model is similar to that of Leitenstorfer and Tutz (2007), who used B-splines and likelihood based boosting. However, they do not enforce $g(0) = 0$, and their estimated dose-response curve is negative for concentrations of SO_2 less than 25 microns. The approach proposed here does not suffer from this problem, and we meet constraints (i) to (iii) by modelling $g(z_{t-k})$ as an (Integrated) I-spline of order 3 (Ramsay (1988)). Specifically, we model

$$g(z_{t-k}) = \sum_{j=1}^{q} I_j(z_{t-k})\beta_j,$$

where $I_j(z_{t-k})$ is the jth I-spline basis function evaluated at z_{t-k}, and $\beta = (\beta_1, \ldots, \beta_q)$ are the associated regression parameters. Integrated spline basis functions are monotone, so that a (strictly) increasing curve is obtained if the associated parameters are constrained to be (positive) non-negative. Therefore our proposed model is given by

$$\begin{aligned} Y_t &\sim \text{Poisson}(\mu_t) \quad \text{for } t = 1, \ldots, n, \\ \ln(\mu_t) &= \mathbf{x}_t^{\mathrm{T}}\alpha + \sum_{j=1}^{q} I_j(z_{t-k})\beta_j, \qquad (2) \\ &\text{Subject to } \mathbf{A}\beta \geq \mathbf{0}, \end{aligned}$$

where $\mathbf{A}$ is the $q \times q$ identity matrix. In this model q denotes the number of I-spline basis functions, and hence controls the smoothness of the estimated dose-response curve. Therefore as the dose-response curve is likely to be smooth, a small value of q should be sufficient. However, the linear inequality constraints that impose monotonicity also smooth the curve, meaning that increasing the number of basis functions does not have a large effect on the resulting estimate. Therefore, the estimated curves should show little difference for small to moderate values of q. We choose q by minimising AIC, although due to the linear inequality constraint, Hughes *et al* (2003) describe that the effective number of parameters cannot be

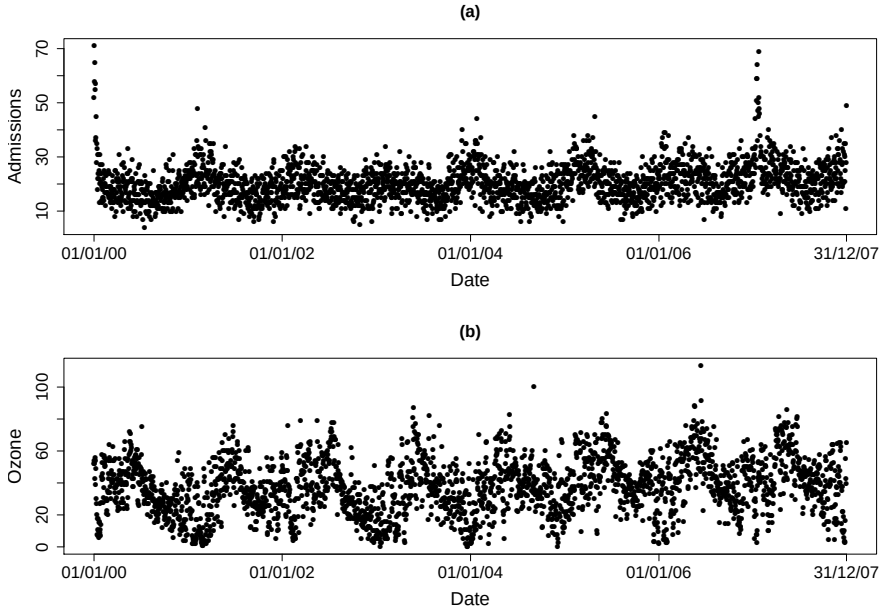


FIGURE 1. Daily counts of (a) respiratory admissions and (b) ozone concentrations for the years 2000 to 2007.

computed exactly. In this paper we approximate this by the total number of parameters minus the number that have been constrained to equal zero by the restriction $\mathbf{A}\beta \geq \mathbf{0}$.

Estimation for this model is implemented using maximum likelihood methods, which involve the solution of a quadratic programming problem with inequality constraints at each stage of the iteratively re-weighted least squares algorithm. Due to the constraints confidence intervals for the dose-response curve cannot be calculated in the usual way, because this results in the lower interval having negative values. Instead a bootstrapping approach is adopted, which is based on re-sampling from the vector of fitted values many times and taking the confidence interval to be the 2.5 and 97.5 percentiles of the estimate from the bootstrapped samples.

3 Glasgow application

3.1 Data

Our response data are daily counts of respiratory related hospital admissions for Glasgow between 2000 and 2007, while our pollutant is daily mean

ozone concentrations, both of which are displayed in Figure 1. In addition, our model also includes minimum temperature levels, a day of the week indicator, an 'epidemic' indicator, and a natural cubic spline of time to model the cyclical and seasonal trends. We included a day of the week indicator as we found that there were higher numbers of respiratory admissions at the weekend compared to the week day, and our 'epidemic' indicator captures the unusually large number of admissions, Figure 1(a), in late 2006 and early 2007.

3.2 Model building

There were a number of days for which the average ozone concentration and/or some of the explanatory variables were missing, so we restricted the analysis to days with no missing values across the full set of predictors. Ozone concentrations were lagged by one day, and 5 I-spline basis functions (i.e. $q = 5$) were chosen as it minimised AIC. A natural cubic spline of time was used to model the trend as it is numerically stable and gives a smooth estimated trend. To determine the number of knots we firstly made visual comparisons of the fit for different values, in order to give us an idea of how many knots to consider. Our final choice of 7 knots per year, 56 in total, was made by minimising AIC.

3.3 Results

We compare our model with one that incorporates an unconstrained dose-response curve, so that the deficiencies in the latter can be observed. We fitted a natural cubic spline of ozone to model the shape of the relationship in the unconstrained model and, to ensure both curves have the same level of smoothness, 5 knots per year were used. Figure 2 displays the risk of hospital admission due to ozone exposure relative to the minimum observed concentration for both models. The unconstrained curve in panel (a) is unrealistic, because the relative risk is below one for ozone concentrations of approximately 1 to 21μgm^3, and after 62μgm^3 the risk of hospital admission decreases. In contrast, the constrained model in panel (b) does not give a relative risk below one for any concentration of ozone, and therefore does not imply it could be beneficial to your health. This curve is also biologically plausible, because increasing ozone concentrations result in increasing health risks.

4 Future Work

In future we will develop a Bayesian implementation of this model, as well as extending it to estimate the average dose-response curve across multiple cities.

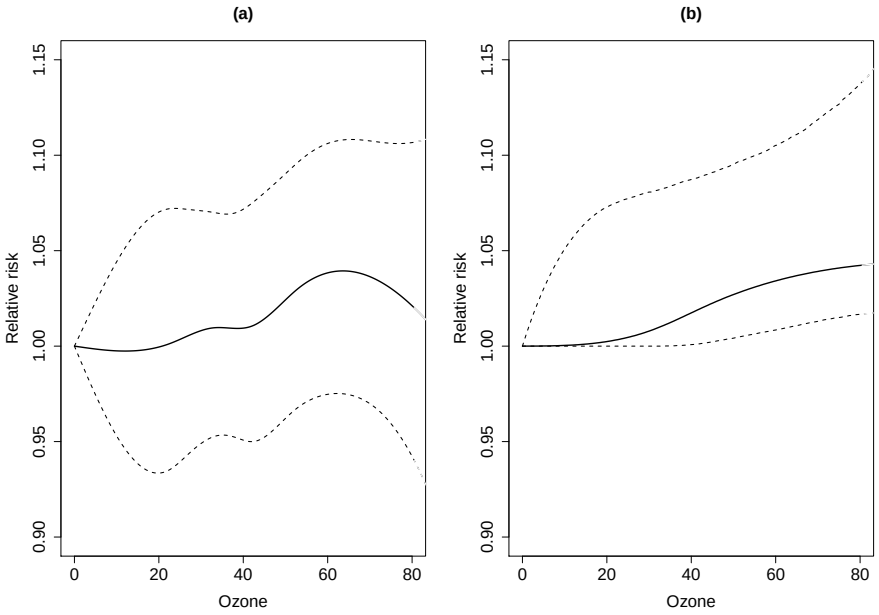


FIGURE 2. Relative risk curves and associated 95% confidence intervals for (a) the constrained model and (b) the unconstrained model.

References

Dominici, F., Daniels, M., Zeger, S., and Samet, J. (2002). Air Pollution and Mortality: Estimating Regional and National Dose-Response Relationships. *Journal of the American Statistical Association*, **97**, 100-111.

Hughes, A., and King, M. (2003). Model selection using AIC in the presence of one-sided information *Journal of Statistical Planning and Inference*, **115**, 397-411.

Leitenstorfer, F., and Tutz, G. (2007). Generalized monotonic regression based on B-splines with an application to air pollution data *Biostatistics*, **8**, 654-673.

Ramsay, J.O. (1988). Monotone Regression Splines in Action *Statistical Science*, **4**, 425-461.

Shaddick, G., Lee, D., Zidek, J.V., and Salway, R. (2008). Estimating Exposure Response Functions using Ambient Pollution Concentrations *The Annals of Applied Statistics*, **2**, 1249-1270.

Correlated GMRF priors for multivariate age-period-cohort models

Andrea Riebler[1], Leonhard Held[1] and Håvard Rue[2]

[1] Biostatistics Unit, Institute of Social and Preventive Medicine, University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland;
Email: {andrea.riebler, leonhard.held}@ifspm.uzh.ch,
[2] Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway; Email: havard.rue@math.ntnu.no

Abstract: Multivariate age-period-cohort models have recently been proposed for the analysis of heterogeneous time trends. For a fully Bayesian analysis, Gaussian Markov random field (GMRF) priors are typically used. However, standard GMRF priors do not account for a potential dependence between outcomes. We present an extended approach based on correlated smoothing priors and correlated overdispersion parameters. Algorithmic routines are based on either Markov chain Monte Carlo or integrated nested Laplace approximations. Results are discussed for data on female mortality in Denmark and Norway and compared by means of DIC, proper scoring rules and the marginal likelihood.

Keywords: Bayesian analysis; Gaussian Markov random field; INLA; Multivariate age-period-cohort model; Uniform correlation matrix.

1 Introduction

Age-period-cohort (APC) models are used to analyse mortality or disease counts stratified by age and period. For the case in which rates are available for multiple health outcomes multivariate APC models have been proposed, see e. g. Jacobsen et al. (2004) or Riebler and Held (2010). A joint analysis may borrow strength from a set of shared effects, for example, the age effects while possibly identifying different period or cohort effects. Within a Bayesian setting, typically, both overdispersion parameters and smoothing priors on the time trends are assumed to be independent across outcomes. Hence, a potential dependence between the outcomes is not captured.
We present an extended approach based on correlated overdispersion parameters and correlated smoothing priors. The latter involves a Kronecker product structure composed of the inverse of a uniform correlation matrix and the precision matrix of the univariate second-order random walk (RW2). Fully Bayesian inference is conducted by either Markov chain Monte Carlo (MCMC) or integrated nested Laplace approximations (INLA) (Rue et al., 2009). The methodology will be applied to mortality rates among

Danish and Norwegian women and models will be compared based on proper scoring rules (Gneiting and Raftery, 2007), the well-known deviance information criterion (DIC) and the marginal likelihood.

2 The correlated multivariate APC model

Let n_{ijs} denote the number of persons under risk in age group i ($i = 1, \ldots, I$), period j ($j = 1, \ldots, J$) and health outcome s ($s = 1, \ldots, S$). We assume that the number of disease cases or deaths y_{ijs} follows a Poisson distribution with mean $n_{ijs}\lambda_{ijs}$, where in the most general formulation

$$\eta_{ijs} = \log(\lambda_{ijs}) = \mu_s + \theta_{is} + \phi_{js} + \psi_{ks}. \tag{1}$$

Here, μ_s is the outcome-specific intercept, and θ_{is}, ϕ_{js} and ψ_{ks} are outcome-specific age, period and cohort effects, respectively. The cohort index k depends on age index i and period index j and is defined as $M \times (I - i) + j$ where M is the ratio of the widths of the age group and period intervals. Simpler models can be obtained, for example by assuming shared period effects. Then, the linear predictor is

$$\eta_{ijs} = \log(\lambda_{ijs}) = \mu_s + \theta_{is} + \phi_{j} + \psi_{ks}. \tag{2}$$

Since we are in a Bayesian context all parameters are treated as random variables and prior distributions need to be assigned. We use a flat prior for each μ_s and assume that second differences of shared time effects, here the period effects, are independent Gaussian variables. For outcome-specific time effects, here the age and cohort effects, we use a correlated GMRF prior with precision matrix $\mathbf{P} = \mathbf{C}^{-1} \otimes \mathbf{R}$. Here, $\mathbf{C}^{-1}$ is the inverse of the $S \times S$ uniform correlation matrix $\mathbf{C} = (1 - \rho)\mathbf{I} + \rho\mathbf{J}$, where ρ denotes the correlation parameter, $\mathbf{I}$ the identity matrix and $\mathbf{J}$ is a matrix of ones, and $\mathbf{R}$ is the precision matrix of the univariate RW2 (see Rue and Held, 2005, page 110). This formulation corresponds to a multivariate RW2 with correlated increments. Note that we assign to each time-scale an individual precision and in the case of outcome-specific effects an individual correlation parameter. Sum-to-zero constraints are assumed for each parameter vector, in (2) θ_s, ϕ and ψ_s with $s = 1, \ldots, S$.

To adjust for unobserved heterogeneity we introduce further outcome-specific variables z_{ijs} into the linear predictor (1). Typically, these overdispersion parameters are assumed to be independent Gaussian variables with mean zero and unknown variance (Besag et al., 1995). We propose correlated overdispersion parameters and set $z_{ij} = (z_{ij1}, \ldots, z_{ijS})^\top \sim \mathrm{N}(0, \tau_z^{-1}\mathbf{C})$ for all i and j, where τ_z denotes the precision of the overdispersion.

All of the up to eight hyperparameters (four precisions and up to four correlations) are treated as unknown. Suitable gamma-hyperpriors are assigned to the precisions. To each correlation ρ we apply Fisher's z-transformation

$$\tilde{\rho} = \log\left(\frac{1 + \rho}{1 - \rho}\right), \quad -1 < \rho < 1,$$

and assign a Gaussian prior with mean zero and variance 0.2^{-1} to $\tilde{\rho}$, corresponding to a U-shaped prior for correlation ρ. To ensure positive definiteness of $\mathbf{C}$ the additional constraint $\rho > -1/(S-1)$ is required.

3 Implementation

Algorithmic routines based on MCMC were implemented in the low-level programming language `C` using the `GMRFLib` library (Rue and Held, 2005). Following Besag et al. (1995), we reparameterised the model from z_{ijs} to η_{ijs} to obtain multivariate normal full conditional distributions for the intercepts and time effects. Block updating allows the proper incorporation of the sum-to-zero constraints for the time effects. For the precisions also Gibbs sampling is used. The vector $\eta_{ij} = (\eta_{ij1}, \ldots, \eta_{ijS})^\top$ has a non-standard distribution. It is updated using multivariate Metropolis-Hastings steps with a GMRF proposal distribution based on a second-order Taylor approximation of the log-likelihood. For the correlation parameters Metropolis-Hastings updates based on a random walk proposal are used, such that acceptance rates around 40% are achieved.
An attractive and fast alternative to MCMC in the class of latent Gaussian random field models is INLA (Rue et al., 2009). This approach directly computes very accurate approximations to the posterior marginal distributions, so that MCMC sampling becomes redundant. We included a new option in the `inla` programme to correlate a wide range of latent GMRF models based on a uniform correlation structure. The methodology can be applied using the `R`-package `INLA` (see `www.r-inla.org`). Here, we use the `INLA` package built on 09.04.2010.

4 Model choice

The DIC is frequently used for model comparison. It is the sum of the posterior saturated deviance $\bar{D}$, a measure for model fit, and the effective number of parameters p_D, a measure for model complexity. Within both MCMC and INLA, estimates for DIC can be calculated. However, for hierarchical models with many random effects, as in (1) with included overdispersion parameters, the use of DIC has recently been criticised (Plummer, 2008).
An alternative are proper scoring rules, e.g. the mean Dawid-Sebastiani score (Riebler and Held, 2010). To account for the correlation potentially present in multiple outcomes and captured by using correlated GMRF priors, this score needs to be adapted. We denote this generalised form as multivariate mean David-Sebastiani score $\overline{\text{MDSS}}$. Within MCMC we used approximate leave-one-block-out cross-validation based on replicating the vector $\eta_{ij} = (\eta_{ij1}, \ldots, \eta_{ijS})^\top$ and subsequently the observation vector $\mathbf{y}_{ij} = (y_{ij1}, \ldots, y_{ijS})^\top$ (Marshall and Spiegelhalter, 2003). These replicated

data points can now be used to calculate the $\overline{\text{MDSS}}$ as:

$$\overline{\text{MDSS}} = \frac{1}{IJ} \sum_{i,j} \left[\left(\mathbf{y}_{ij} - \overline{\mathbf{y}_{ij}^{\text{rep}}} \right)^\top \{\boldsymbol{\Sigma}_{ij}^{\text{rep}}\}^{-1} \left(\mathbf{y}_{ij} - \overline{\mathbf{y}_{ij}^{\text{rep}}} \right) + \log \left| \boldsymbol{\Sigma}_{ij}^{\text{rep}} \right| \right]$$

where $\overline{\mathbf{y}_{ij}^{\text{rep}}} = (\overline{y_{ij1}^{\text{rep}}}, \ldots, \overline{y_{ijS}^{\text{rep}}})^\top$ and $\overline{y_{ijs}^{\text{rep}}}$ is the mean of the N replicated observation samples $\mathbf{y}_{ijs}^{\text{rep}} = (y_{ijs_{(1)}}^{\text{rep}}, \ldots, y_{ijs_{(N)}}^{\text{rep}})^\top$. Analogously, $\boldsymbol{\Sigma}_{ij}^{\text{rep}}$ represents the empirical covariance matrix of $(\mathbf{y}_{ij1}^{\text{rep}}, \ldots, \mathbf{y}_{ijS}^{\text{rep}})^\top$.
Furthermore, INLA returns an estimate of the log marginal likelihood $\log(p(\mathbf{y}))$. Usually the marginal likelihood is difficult to use for hierarchical GMRF models in which the prior is improper (here because of the RW2). However, for comparing models that only differ by the inclusion of correlation but have the same underlying first-level structure, e.g. (2), $\log(p(\mathbf{y}))$ can be used for model choice.

5 Mortality of Danish and Norwegian women

We analyse data on overall mortality, aggregated to 5-year age group and period intervals (i.e. $M = 1$), for all Danish and Norwegian women aged 0-84 years during the period 1960-1999 (Jacobsen et al., 2004). In an uncorrelated multivariate APC analysis Riebler and Held (2010) classified the aPc_z model with separate age and cohort effects but joint period effects as best. Here, we compare the aPc_z model with independent RW2 priors for θ_s, ϕ and ψ_s, $s = 1, 2$, to three different correlated models. Either age and cohort effects ($\text{a}^\star\text{Pc}_z^\star$ model), or the overdispersion parameters ($\text{aPc}_{z^\star}$ model) are correlated. Both correlated time and overdispersion parameters are specified in model $\text{a}^\star\text{Pc}_{z^\star}^\star$. For all models MCMC and INLA produce virtually identical results. The posterior correlation estimates of the $\text{a}^\star\text{Pc}_{z^\star}^\star$ model clearly indicate the dependence present between the outcomes, compare Figure 1. Table 1 shows the model choice criteria obtained by MCMC and INLA for all models. The $\text{a}^\star\text{Pc}_{z^\star}^\star$ model is clearly preferred.
Figure 2 shows estimates of relative risks for the $\text{a}^\star\text{Pc}_{z^\star}^\star$ model together with estimates of the uncorrelated aPc_z model. The estimates of the correlated model are smoother and the credible regions are smaller. For some interpretation of the relative risks see Riebler and Held (2010).

6 Conclusion

We proposed the use of correlated GMRF priors for multivariate age-period-cohort models and implemented these models based on a uniform correlation structure in MCMC and INLA. We illustrated the methodology on female mortality in Norway and Denmark and received virtually identical results with both approaches, MCMC and INLA. A correlated

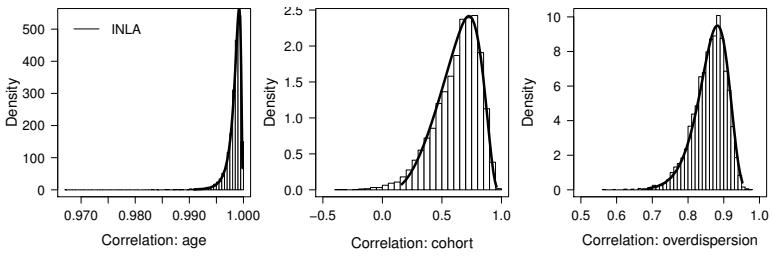


FIGURE 1. Posterior correlation estimates of the a$\star$Pc$\star_{z^\star}$ model. Approximated posterior marginals of INLA and corresponding histograms based on 5000 MCMC samples after 20 000 burn-in iterations and a thinning of 20 are shown.

TABLE 1. Model choice criteria obtained from MCMC and INLA. For both approaches deviance summaries are given. In addition, the multivariate mean Dawid-Sebastiani score $\overline{\text{MDSS}}$ and the log marginal likelihood are shown. (For each measure the best value is indicated in bold.)

	aPc_z	$\text{a}^\star\text{Pc}^\star_z$	$\text{aPc}_{z^\star}$	$\text{a}^\star\text{Pc}^\star_{z^\star}$
MCMC model choice				
$\overline{\text{D}}$	301.2	304.4	293.5	**292.4**
p_D	201.1	194.4	183.6	**176.4**
DIC	502.2	498.8	477.1	**468.8**
$\overline{\text{MDSS}}$	19.79	19.71	19.40	**19.39**
INLA model choice				
$\overline{\text{D}}$	301.6	304.6	293.6	**292.5**
p_D	201.1	194.2	183.7	**176.3**
DIC	502.7	498.9	477.3	**468.8**
$\log(p(\mathbf{y}))$	-1799.6	-1765.6	-1776.3	$\mathbf{-1741.1}$

model structure lead to more precise relative risk estimates and was clearly preferred in this application.

However, benefits of correlated multivariate APC models might not only be in terms of model choice criteria and the improved precision of relative risk estimates. When projecting e.g. mortality rates of one health outcome a correlated joint analysis with a set of comparable outcomes may borrow strength from these and thus lead to more accurate projections.

Acknowledgments: This work received support from the Swiss National Science Foundation and the Research Council of Norway.

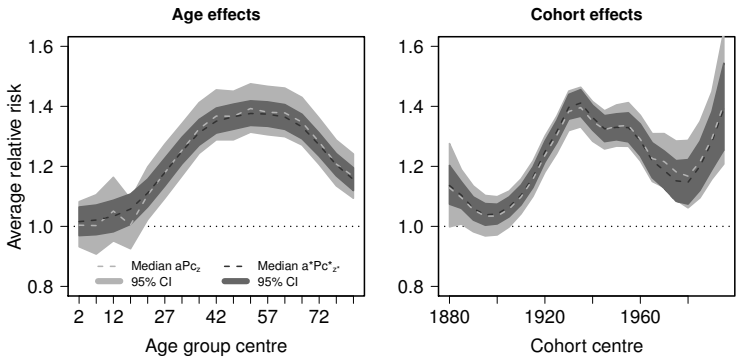


FIGURE 2. Average relative risk of death for Danish compared with Norwegian women analysed by the aPc_z and $\mathrm{a}^\star\mathrm{Pc}^\star_{z^\star}$ model.

References

Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science*, **10**, 3-41.

Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359-378.

Jacobsen, R., von Euler, M., Osler, M., Lynge, E. and Keiding, N. (2004). Women's death in Scandinavia - What makes Denmark different? *European Journal of Epidemiology*, **19**, 117-121.

Marshall, E.C. and Spiegelhalter, D.J. (2003). Approximate cross-validatory predictive checks in disease-mapping methods. *Statistics in Medicine*, **22**, 1649-1660.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics*, **9**, 523-539.

Riebler, A., and Held, L. (2010). The analysis of heterogeneous time trends in multivariate age-period-cohort models. *Biostatistics*, **11**, 57-69.

Rue, H., and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. London: Chapman & Hall/CRC Press.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, **71**, 319-392.

Efficient semi-parametric SNP genotyping

Ralph C.A. Rippe[1], Paul H.C. Eilers[2]

[1] Institute of Psychology, Leiden University, P.O. Box 9555 2300 RB Leiden, The Netherlands (RRippe@fsw.leidenuniv.nl)
[2] Department of Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands

Abstract: The genotypes of single nucleotide polymorphisms (SNPs) can be determined by modeling clusters in a scatter plot of fluorescence signals. We estimate mixtures with three semi-parametric log-concave densities, based on tensor product P-splines.

Keywords: Log-concave, density estimation, P-splines, tensor product.

1 Introduction

The three panels in Figure 1 illustrate an important statistical challenge in modern genetics. The variables a and b in the left panel represent fluorescence signals. They have been measured on microarrays, designed for large-scale genotyping of single nucleotide polymorphisms (SNP, pronounced as snip). Each SNP can have two different forms, so-called alleles, which we indicate by A and B. Because chromosomes in healthy human cells form pairs, each SNP manifests itself as an unordered pair AA, AB or BB, called the genotype. In theory, signal a (b) is proportional to the number of A (B) alleles. In practice this is not the case, because we do not see three tight groups of dots, but elongated clusters. However, visually it seems quite clear which cluster represents which genotype (counting clockwise: BB, AB and AA). The challenge is to assign each dot in the plot to the most probable cluster, and hence determine the genotype.

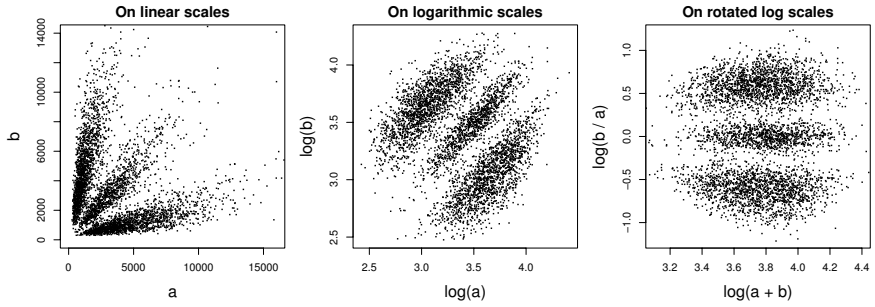


FIGURE 1. Data (Affymetrix 50k) transformation for signals on chromosome 1. Signal a (b) represents allele A (B).

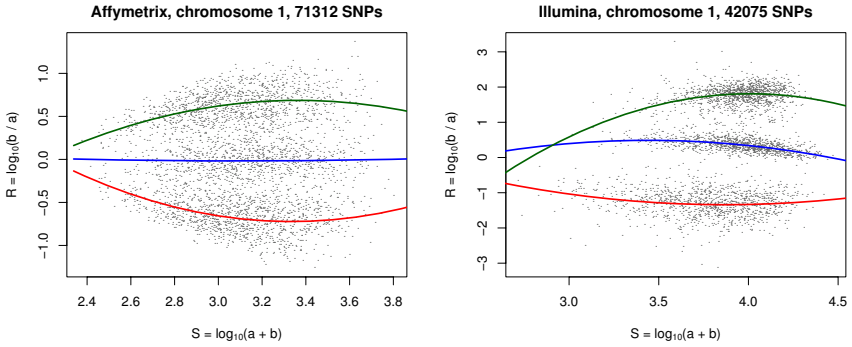


FIGURE 2. Parametric model on Affymetrix array (left) and IIlumina array (right). A random selection of 3500 SNPs is plotted as dots.

In earlier work (Rippe et al., 2009) we proposed a parametric model to determine the genotypes using the transformed data $S = \log(a + b)$ and $R = \log(b/a)$; logarithms are to base 10. The model assumed a mixture of three linear or quadratic curves, to which normally distributed noise is added. The curve parameters and the SD of the noise are specific to each cluster. From these (and the overall probabilities of the clusters) follow three membership probabilities for each SNP; the largest probability determines the genotype.

For Affymetrix microarrays this model (especially the quadratic form) describes the data quite well (left panel in Figure 2). The upper and lower clusters have about the same size and the quadratic trends do not interfere with eachother. We see a different pattern for Illumina Infinium arrays. The clusters are far from symmetric and the quadratic trend lines intersect, as the right panel of Figure 2 shows. Also we encountered data that seemed to violate the assumption of constant noise within a clusters. This was our motivation for developing a model based on mixtures of log-concave semi-parametric densities.

2 Semi-parametric log-concave densities

Eilers and Marx (2007) described how to use P-splines (Eilers & Marx, 1996) to obtain smooth multidimensional density estimates. First we discuss the one-dimensional case.

For one dimension, let y_i denote the count in bin i of a histogram and let u_i be the bin midpoint, with $i = 1, \ldots, n$. The vector of counts is denoted

by $\boldsymbol{y} = y_i$. We model the expected values of the counts as

$$\begin{aligned} \mu_i &= E(y_i) = \exp\left(\sum_{j=1}^{c} b_j(u_i)\theta_j\right) \quad \text{or} \\ \boldsymbol{\mu} &= \boldsymbol{B\theta}, \end{aligned} \tag{1}$$

where $\boldsymbol{B} = [b_j(u_i)]$ is a $(n \times c)$ B-spline basis, with c relatively large. Assuming Poisson distributed counts, we optimize the penalized log-likelihood

$$l^* = \sum_{i=1}^{n}(y_i \log \mu_i - \mu_i) - \lambda \sum_{j=1}^{c}(\Delta^3 \theta_j)^2/2. \tag{2}$$

Smoothness is tuned with the parameter λ. Notice that we use third order differences. This has the effect that for larger values of λ the vector $\boldsymbol{\theta}$ tends towards a quadratic series, because for such a series third order differences vanish and the penalty is zero. Unless the series of counts $\boldsymbol{y}$ has a J, U, or L shape, $\boldsymbol{\theta}$ will approach a mountain parabola and the estimated density will show a unimodal log-concave shape. This is a desirable property for components of the mixtures we consider.
Setting the derivative of l^* to zero gives

$$\boldsymbol{B}'(\boldsymbol{y} - \boldsymbol{\mu}) = \lambda \boldsymbol{D}'\boldsymbol{D\theta}, \tag{3}$$

where $\boldsymbol{D}$ is a matrix of contrasts such that $\boldsymbol{D\theta} = \Delta^3\boldsymbol{\theta}$. Linearization of (3) leads to

$$(\boldsymbol{B}'\tilde{\boldsymbol{W}}\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D})\boldsymbol{\theta} = \boldsymbol{B}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}}, \tag{4}$$

where $\boldsymbol{z} = \boldsymbol{\eta} + \tilde{\boldsymbol{W}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$ is the working variable and $\boldsymbol{\eta} = \boldsymbol{B\theta}$. The matrix $\boldsymbol{W} = \text{diag}(\boldsymbol{\mu})$ and $\tilde{\boldsymbol{\theta}}$, $\tilde{\boldsymbol{\mu}}$ are approximations to the solution of (4). This system is iteratively solved until convergence.
In two dimensions we use the same idea, but now a two-dimensional histogram is formed, and the log of the density is formed by a sum of tensor products of B-splines. We sketch the adaptations that have to be made. Let $\boldsymbol{Y} = \{y_{ih}\}$ be an n_1 by n_2 matrix of counts in a two-dimensional $n_1 \times n_2$ histogram. The center of bin (i, h) is given by (u_i, v_h). The expected values are modeled using tensor product B-splines. Two bases are computed, $\boldsymbol{B}$, with c_1 columns, based on $\boldsymbol{u}$ and $\breve{\boldsymbol{B}}$, with c_2 columns, based on $\boldsymbol{v}$. From these follows a c_1 by c_2 matrix $\boldsymbol{\Theta}$ of coefficients and the matrix of expected values is computed as

$$\boldsymbol{M} = \exp(\boldsymbol{B\Theta}\breve{\boldsymbol{B}}'). \tag{5}$$

Like in the one-dimensional case, a penalized Poisson log-likelihood is optimized. The penalty is more complex, because both rows and columns of $\boldsymbol{\Theta}$ are penalized. If $||\boldsymbol{X}||_F$ indicates the Frobenius norm of the matrix $\boldsymbol{X}$, the sum of the squares of its elements, the penalty is

$$\text{Pen} = \lambda||\boldsymbol{D\Theta}||_F/2 + \breve{\lambda}||\boldsymbol{\Theta}\breve{\boldsymbol{D}}'||_F/2, \tag{6}$$

where $\boldsymbol{D}$ and $\breve{\boldsymbol{D}}$ are matrices of the proper dimensions ($c_1 - 3$ by c_1 and $c_2 - 3$ by c_2) that form third differences.
One could vectorize $\boldsymbol{Y}$, $\boldsymbol{M}$ and $\boldsymbol{\Theta}$ and form the Kronecker product of $\breve{\boldsymbol{B}}$ and $\boldsymbol{B}$ to mold the equations into the familiar GLM matrix-vector shape. The penalty also has to be exploded to fit this format. It is however very inefficient to do so. Instead, the fast array algorithm can be used (Currie et al. 2006), leading to enormous savings in computation time and memory use. The details are a bit involved; we skip them here.
Once a (two-dimensional) histogram smoother as described above is available, mixtures can be estimated with the EM algorithm. Eilers & Borgdorf (2007) illustrated this for a one-dimensional log-concave mixture, but their approach can easily be extended to two dimensions. If approximations to the mixture components are estimated, the counts in the histogram bins can be split proportionally into pseudo-counts, giving three pseudo-histograms, one for each component; this is the E step. From each pseudo-histogram a smooth density is estimated, giving the next approximation; this the M step. The splitting and fitting is repeated until convergence.
To start the iterations the data are split by two horizontal lines. The splitting levels can be inferred visually from one array and then applied to similar samples. For Affymetrix we used -0.2 and 0.2, for Illumina -0.6 and 0.6. These values are not critical.
The speed of the EM algorithm depends on the separation of the components. In the present example it is quite good, so 10 iterations or so are enough. We use 13 times 13 tensor products of B-splines and a histogram with 100 by 100 bins. The total computation time is of the order of 10 seconds.
Interestingly, once the initial two-dimensional histogram is formed, computation time no longer depends on the original number of SNPs, which can be over 10^6 for modern microarrays.

3 Results

Figure 3 shows the semi-parametric mixture results for an Affymetrix array, and Figure 4 shows the mixture results for an Illumina array, as opposed to the parametric estimations in Figure 2. Comparing the two models within one platform, the results are highly similar, as expressed in the distribution of the maximum cluster membership probabilities, p_{max}, shown in Figure 5. For over 95% of the SNPs the probability exceeds 0.8, indicating reliable genotyping. The same Figure (5) also shows that the Illumina platform provides a clearer division (a stronger L-shape) using both types of models. Furthermore, for the latter platform, the semi-parametric model provides fewer low probabilities compared to the parametric model.
Furthermore, we found that varying the number of bins ($n = 50, \ldots, 200$) didn't influence the shape and contents of the final density components.

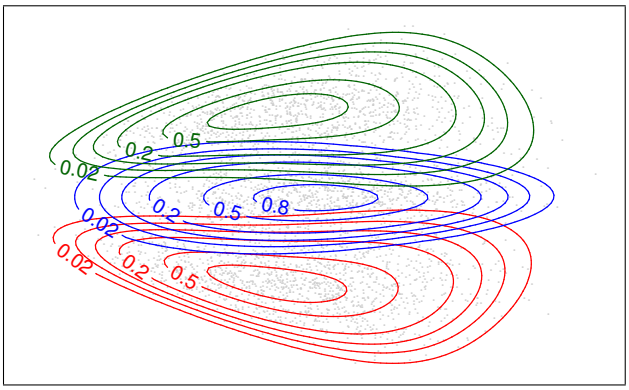


FIGURE 3. Observations and contour lines of semi-parametric mixture components, based on an *Affymetrix* array, chromosome 1. Normalized contours (mode set to 1) are shown at [0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 0.8].

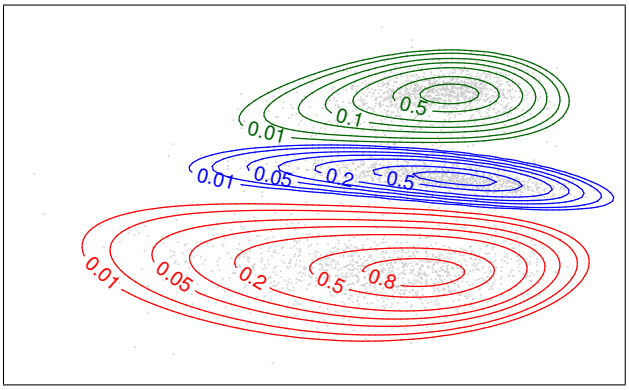


FIGURE 4. Observations and contour lines of semi-parametric mixture components, based on an *Illumina* array, chromosome 1. Normalized contours (mode set to 1) are shown at [0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 0.8].

Also, the number of B-spline bases used in the tensor product only affected computation times, but not the estimation results.

4 Discussion

Based on the results in Figure 5, we argue that, despite the asymmetric shapes, the Illumina platform gives a better classification compared to Affymetrix, even when using a parametric model. Furthermore, the difference between the parametric and semi- parametric models within the same platform is relatively small. However, the semi-parametric model can handle more complex situations (automatically).

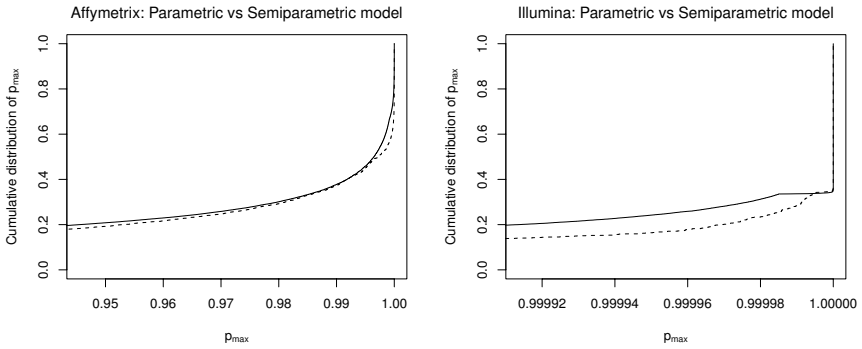


FIGURE 5. Cumulative distributions of the largest membership probabilities, based on all SNPs in the array. The left panel shows parametric (solid line) and semi- parametric probabilities (dotted line) for the Affymetrix array shown in Figure 3. The right panel shows probabilities for the Illumina array shown in Figure 4. In both panels, the horizontal axis starts where the cumulative probability distribution equals 0.2.

Finally, additional comparisons can provide more insight into the effectiveness of the approach proposed above. It is (very) interesting to see how the semi-parametric single array genotype calls for Illumina relate to the BeadStudio results; the latter software uses multiple arrays simultaneously to call the genotype for a single SNP. This is however beyond the scope of this paper and will be discussed elsewhere.

References

Currie, I.D., Durban, M. and Eilers, P.H.C. (2006). Generalized linear array models with applications to multidimensional smoothing. *J. R. Statist. Soc. B* **68**, 259–280.

Eilers, P.H.C. and Borgdorff, M.W. (2007). Non-parametric log-concave , mixtures. *Computational Statistics & Data Analysis*. **51**, 5444 5451.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing using B-splines and penalized likelihood (with Comments and Rejoinder). *Statistical Science*, **11**(2), 89-121.

Eilers, P.H.C. and Marx, B.D. (2007). Multidimensional Density Smoothing with P-splines. *Proceedings of the 23rd International Workshop on Statistical Modelling.*

Rippe, R.C.A., French, P.J., Eilers, P.H.C. & Meulman, J.J. (2009)., Improved SNP genotyping using model-based calibration. *Proceedings of the 24th International Workshop on Statistical Modelling.*

A new flexible direct ROC regression model: Detection of cardiovascular risk factors by anthropometry.

M. X. Rodríguez-Álvarez[13], J. Roca-Pardiñas[2], C. Cadarso-Suárez[13]

[1] Dept. of Statistics and Operations Research, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain. mariajose.rodriguez.alvarez@usc.es.
[2] Dept. of Statistics and Operational Research. University of Vigo, 36208 Vigo, Spain.
[3] Instituto de Investigació Sanitaria de Santiago (IDIS), Santiago de Compostela, Spain.

Abstract: In this work, a new estimator for the conditional ROC curve, based on direct methodology, is presented. In this approach, the effect of the covariates and false positive fraction on the ROC curve is modelled non-parametrically using generalized additive models (GAM) combined with local polynomial kernel smoothers. Our method allows for incorporation of more than one covariate in the regression model for the ROC curve and the possible interaction between them. The proposed model's performance is examined by means of simulations. Lastly, endocrine data are analysed with the aim of assessing the performance of Body Mass Index (BMI) in predicting clusters of cardiovascular disease (CVD) risk factors in an adult population in Galicia (NW Spain), with adjustment for age and gender.

Keywords: ROC curve; generalized additive models; local polynomial kernel smoothing; anthropometric measures.

1 Introduction

The discriminatory capacity of a continuous diagnostic test Y in distinguishing between diseased (D) and healthy ($\bar{D}$) subjects, is usually measured by means of the receiver operating characteristic (ROC) curve (Metz, 1978). In many practical situations, however, a marker's discriminatory capacity may be affected by a set of covariates X. The ROC curve may thus be of little value if important covariates are omitted. Moreover, in such situations, interest must be focused on assessing marker Y's discriminatory capacity by reference to the values assumed by $\mathbf{X}$. If the conditional survival functions of Y_D and $Y_{\bar{D}}$, given $\mathbf{X}$, are denoted $S_{D\mathbf{X}}$ and $S_{\bar{D}\mathbf{X}}$ respectively, the conditional ROC curve is defined as

$$ROC_{\mathbf{X}}(t) = S_{D\mathbf{X}}\left(S^{-1}_{\bar{D}\mathbf{X}}(t)\right), t \in (0,1).$$

To study the effect of covariates on the accuracy of a diagnostic test, various ROC regression methodologies have been proposed. This work is focused on direct methodology (see e.g. Alonzo and Pepe, 2002). More precisely, the main goal of this paper is to present a new flexible estimator of the conditional ROC curve based on the ROC-GAM regression model given by

$$ROC_{\mathbf{X}}(t) = g\left(\alpha + \sum_{k=1}^{p} f_k(X_k) + h(t)\right), t \in (0,1), \tag{1}$$

where $f_j(\cdot)$ and $h(\cdot)$ are assumed to be smooth and unknown functions. It should be noted that, in order to guarantee the identification of model (1), we introduce a constant α into the model and require a zero mean for the partial functions, $E\{f_j(X_j)\} = 0, j = 1, \ldots, p$ and $E\{h(t)\} = 0$.

2 Estimation procedure

Let $\left\{\left(y_i^{\bar{D}}, \mathbf{x}_i^{\bar{D}}\right)\right\}_{i=1}^{n_{\bar{D}}}$ and $\left\{\left(y_j^{D}, \mathbf{x}_j^{D}\right)\right\}_{j=1}^{n_D}$ be two independent random samples drawn from the healthy and diseased populations, respectively. The key idea for fitting the model (1) is based on the placement values (Hanley and Hajian-Tilaki, 1997) of Y_D defined as $PV_D \equiv S_{\bar{D}X}(Y_D)$. Given that

$$E[I[PV_D \le t] | X] = ROC_X(t),$$

the conditional ROC curve can be viewed as the conditional expectation of the binary variable $B_{Dt} = I[PV_D \le t]$. The ROC-GAM regression model (1) can therefore be viewed as a regression model for B_{Dt}. This suggests (Alonzo and Pepe, 2002) that estimation of the ROC-GAM regression model (1) can be based on the following algorithm:

1. choose a set $T = \{t_l, l = 1, \ldots, n_T\}$ of FPFs;
2. estimate $S_{\bar{D}X}$ on the basis of $\left\{\left(y_i^{\bar{D}}, \mathbf{x}_i^{\bar{D}}\right)\right\}_{i=1}^{n_{\bar{D}}}$;
3. calculate the estimated placement value $PV_j = \hat{S}_{\bar{D}\mathbf{x}_j^D}\left(y_j^D\right), j = 1, \ldots, n_D$, and the binary placement value indicator $\hat{B}_{jt_l} = I[PV_j \le t_l]$, $j = 1, \ldots, n_D,\ l = 1, \ldots, n_T$; and
4. fit the following ROC-GAM binary regression model

$$ROC_X(t) = g\left(\alpha + \sum_{k=1}^{p} f_k(X_k) + h(t)\right), \tag{2}$$

to the data $\left\{\left(\hat{B}_{jt_l}, \left\{\mathbf{x}_j^D, t_l\right\}\right), l = 1, \ldots, n_T, j = 1, \ldots, n_D\right\}$, and obtain the estimates $\widehat{ROC}_X(t)$.

To estimate the binary GAM (2), use is made of the local scoring algorithm with an inner backfitting loop (Hastie and Tibshirani, 1990), based on local linear kernel smoothers (Fan and Gijbels, 1996). Among the advantages of using such smoothers is the possible use of binning type acceleration techniques (Fan, 1994) to reduce computational time and so ensure that the problem can be adequately addressed in practical situations.
It should be noted that the observations used to fit the binary GAM (2) are no longer independent, since each disease observation is 'compared' with all $t_l \in T$. It is well known that in the presence of correlated errors, standard bandwidths selectors fail to work and can result in an over (or under) fit (see e.g. Opsomer et al., 2001). In this study, the cross-validation criterion was used for the automatic choice of bandwidths. As pointed out, this choice of bandwidths may be far from optimal. Nevertheless, the estimation procedure seems to perform reasonably well in the simulation studies presented in Section 3 below.
To implement the estimation procedure presented at the beginning of this section, the conditional survival function in healthy subjects must be estimated, $S_{\bar{D}\mathbf{X}}$ (see Step 2). In this paper, we propose to model the effect of covariates on $Y_{\bar{D}}$ by a non-parametric location-scale regression model, such that

$$Y_{\bar{D}} = \mu_{\bar{D}}(\mathbf{X}) + \sigma_{\bar{D}}(\mathbf{X})\varepsilon_{\bar{D}},$$

where $\mu_{\bar{D}}$ and $\sigma^2_{\bar{D}}$ are the regression and the variance functions respectively, and error $\varepsilon_{\bar{D}}$ is assumed to be independent of the covariates $\mathbf{X}$, with zero mean, unit variance and survival function S. With this configuration, it can be shown that

$$S_{\bar{D}\mathbf{X}}(y) = S\left(\frac{y - \mu_{\bar{D}}(\mathbf{X})}{\sigma_{\bar{D}}(\mathbf{X})}\right).$$

3 Simulation Study

This section reports the results of a simulation study conducted to study the practical behavior of the estimation procedure described in Section 2 above.
Data were simulated from two scenarios, namely,

- Scenario I

$$Y_D = .5\sin(\pi(X_1 + 1)) + .5\exp(X_1) - X_2^2 + .5\varepsilon_D,$$

$$Y_{\bar{D}} = .5\exp(X_1) - 2X_2^2 + .5\varepsilon_{\bar{D}},$$

- Scenario II

$$Y_D = (.5\sin(\pi(X_1 + 1)) + .5\exp(X_1))\,Z + 3X_1(1 - Z) + .5\varepsilon_D,$$

$$Y_{\bar{D}} = .5\exp(X_1)Z + 2X_1(1 - Z) + .5\varepsilon_{\bar{D}},$$

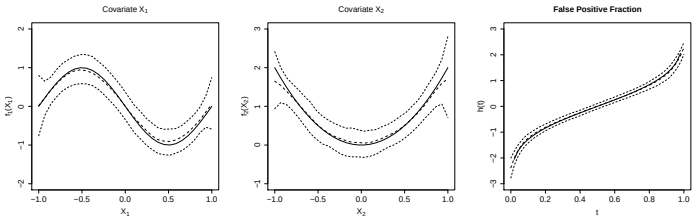


FIGURE 1. Simulation results based on 1000 replicated samples for Scenario I. From left to right: true curves f_1, f_2, and h (solid lines) and average estimates $\hat{f}_1$, $\hat{f}_2$, and $\hat{h}$ (dashed lines). In all cases, the 2.5 and 97.5 simulation quantiles have also been plotted.

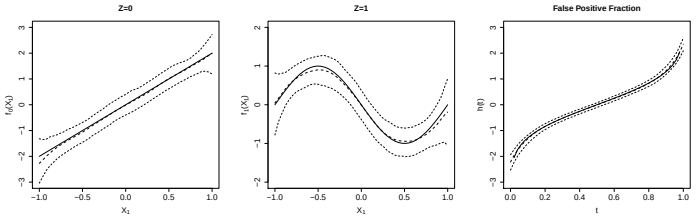


FIGURE 2. Simulation results based on 1000 replicated samples for Scenario II. From left to right: true curves f_0, f_1, and h (solid lines) and average estimates $\hat{f}_0$, $\hat{f}_1$, and $\hat{h}$ (dashed lines). In all cases, the 2.5 and 97.5 simulation quantiles have also been plotted.

where X_1 and $X_2 \sim U[-1,1]$, $Z \sim Be(0.5)$, $\varepsilon_{\bar{D}}$ and $\varepsilon_D \sim N(0,1)$, and Φ denotes the CDF of a standard normal variable. With the above configurations, the corresponding covariate-specific $ROC_X(t)$ are respectively

- I: $ROC_X(t) = \Phi\left(\sin\left(\pi(X_1+1)\right) + 2X_2^2 + \Phi^{-1}(t)\right)$,
- II: $ROC_X(t) = \Phi\left(\left(\sin\left(\pi(X_1+1)\right)\right)Z + 2X_1(1-Z) + \Phi^{-1}(t)\right)$.

Averages of the results are graphically depicted in Figures 1 and 2. The same sample size was considered for both healthy and diseased subjects, with $n = n_D = n_{\bar{D}} = 400$. The good performance of the resulting estimates is evident, with the functional forms of the corresponding true curves being recovered very successfully.

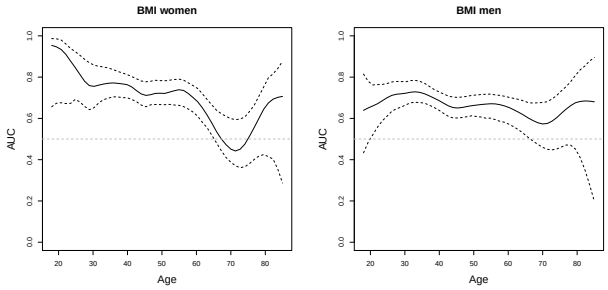


FIGURE 3. AUCs adjusted by age and gender with 95% bootstrap confidence bands for Women and Men.

4 Detecting cardiovascular risk factors by means of anthropometry

We applied the proposed ROC-GAM methodology to an endocrine study, with the aim of assessing the performance of BMI for predicting clusters of cardiovascular risk factors in an age- and gender-adjusted adult population in Galicia (NW Spain). Since it is well established that anthropometric measures perform differently according to gender, the age-by-gender interaction was included in the ROC-GAM regression models.

The study was carried out with a random sample of Galician adult population (2945 subjects, 46.2% men; age range 18-85 years). Subjects having two or more cardiovascular disease risk factors (raised triglycerides, blood pressure and plasma glucose, and reduced HDL-cholesterol) were considered as diseased.

The following multivariate ROC-GAM interaction model was considered:

$$\begin{aligned} ROC_{(Age,Gender)}(t) &= \Phi(\alpha_0 + \alpha_1 \mathbf{1}_{\{Gender=Men\}} \\ &+ f_{Men}(Age)\mathbf{1}_{\{Gender=Men\}} \\ &+ f_{Women}(Age)\mathbf{1}_{\{Gender=Women\}} + h(t)), \end{aligned}$$

where f_{Men} and f_{Women} are smooth functions of Age in men and women, respectively, $h(t)$ is a smooth function of the false positive fraction, and $\mathbf{1}_A$ denotes the indicator function of event A.

Figure 3 depicts the conditional areas under the ROC curve (AUCs) together with the corresponding 95% bootstrap-confidence intervals. In the case of men, the accuracy of the BMI tends to decline progressively and eventually starting to lose significance around age 65 years. With respect to women, the AUC indicate very good discriminatory capacity for the youngest women, with values greater than 0.8.

Acknowledgments: The authors would like to express their gratitude for the support received in the form of the Spanish MEC Grant MTM2008-01603 and Galician Regional Authority (Xunta de Galicia) projects INCITE08PXIB208113PR and PGIDIT07PXIB300191PR, and would also like to thank the Galician Endocrinology & Nutrition Foundation (*Fundacion de Endocrinoloxia e Nutricion Galega - FENGA*) for having supplied the database used in this study

References

Alonzo, T.A. and Pepe, M.S. (2002). Distribution-free ROC analysis using binary regression techniques. *Biostatistics*, **3**, 421-432

Fan, J., and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications*. Chapman & Hall: CRC.

Fan, J. (1994) Fast implementation of non-parametric curve estimators. *Journal of Computational and Graphical Statistics*, **3**, 35-56.

Hanley, J.A. and Hajian-Tilaki, K.O. (1997) Sampling variability of nonparametric estimates of the area under receiver operating characteristic curves: an update. *Academic Radiology*, **4**, 49-58.

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

Metz, C.E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, **8**, 183-298.

Opsomer, J., Wang, Y., Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science, **16**, 134-153.*

Comparing treatment policies in early epilepsy through the joint modelling of pre-randomisation event rates and multiple post-randomisation survival times with extensions

J.K. Rogers[1], J.L. Hutton[1]

[1] Department of Statistics, University of Warwick, Coventry, England, CV4 7AL, J.K.Rogers@warwick.ac.uk

Abstract: We present a simple model that allows a pre-randomisation seizure rate to be jointly modelled with post-randomisation times to first and second seizure. This paper then goes beyond the simple model and considers extensions motivated by the epilepsy data considered. We allow each of the post-randomisation survival times to have different distributions and incorporate a cure fraction.

1 Background

We consider the analysis of data from the MRC Multicentre Trial for Early Epilepsy and Single Seizures (MESS), which was undertaken to assess the differences between two policies: immediate, or deferred treatment in those patients who have had only a single seizure or are in early epilepsy. The question of when to commence treatment with antiepileptic drugs (AEDs) is an area of uncertainty as AEDs frequently come with unpleasant side effects, which can include weight loss or weight gain, altered mood, drowsiness, hair loss, or even polycystic ovarian disease and teratogenicity. For most epilepsy sufferers, the benefits of AEDs will far outweigh the associated risks, but for those individuals who have had only a single seizure, or have mild and infrequent seizures, treatment with AEDs may not be immediately necessary.

MESS randomised 1443 individuals in the early stages of epilepsy to either immediate or deferred treatment. The eligibility criteria for inclusion in the trial were being aged at least one month, having a suitably documented history of at least one clinically definite and unprovoked epileptic seizure and there being uncertainty in the patient and clinician as to whether treatment with AEDs should commence. For those allocated to immediate treatment, subsequent choices of drug and dose were in line with the clinicians' usual practice, whilst for those allocated to the deferred treatment group, treatment with AEDs was withheld until absolutely necessary. Baseline characteristics collected at randomisation included age, sex and information on patients' pre-randomisation seizure frequency

and seizure type. An electroencephalogram (EEG) was also requested for each individual.
Exploratory analysis was carried out on 1425 individuals; 18 were removed due to missing information, assumed missing completely at random. A further five individuals with incomplete information on their pre-randomisation seizure history were excluded from the statistical modelling. Statistical analyses were by intention to treat; interest lay in the treatment policy to which an individual was assigned rather than whether an individual was on treatment at the time of future seizures.

2 Current Model

The MESS data are pre-randomisation seizure counts and times to first and second seizure post-randomisation. In studies of recurrent events, like epilepsy, it is usually only time to first seizure that is considered and yet epilepsy is characterised by multiple seizures, not a single, isolated event. Standard survival analysis would treat the pre-randomisation event count information as a covariate, possibly measuring this quantity as a covariate with error. As an alternative, we have developed methodology that allows individuals' pre-randomisation seizure rate and post-randomisation times to first and second seizure to be jointly modelled, assuming that both outcomes are predicted by (unobserved) seizure rates. We assume that the pre-randomisation event count over a period u_i for individual i, X_i, follows a Poisson distribution with rate $\lambda_i\nu_i$. The parameter λ_i relates to the baseline covariates with additional heterogeneity in the population being modelled through ν_i, assumed to follow a Gamma distribution with expectation 1 and variance $1/\alpha$. Smaller values of α are indicative of higher levels of heterogeneity. Let T_{1i} and T_{2i} be the times from randomisation to first and second seizures respectively and set $Y_{1i} = T_{1i}$ and $Y_{2i} = T_{2i} - T_{1i}$, so that Y_{1i} is the time to first seizure and Y_{2i} is the time from first seizure to the second. Both Y_{1i} and Y_{2i} will be independent and Exponentially distributed with rate $\lambda_i\psi_i\nu_i$, where ψ_i is related to the treatment in some way. In summary, the joint model is specified by the following equations:

$$\begin{aligned} f_{X|\nu}(x_i \mid \nu_i; \lambda_i, u_i) &= \frac{(\lambda_i u_i \nu_i)^{x_i} \exp(-\lambda_i u_i \nu_i)}{x_i!}, \\ f_{Y_1,Y_2|\nu}(y_{1i}, y_{2i} \mid \nu_i; \lambda_i, \psi_i) &= (\lambda_i \psi_i \nu_i)^2 \exp(-\lambda_i \psi_i \nu_i (y_{1i} - y_{2i})), \\ g_\nu(\nu_i; \alpha) &= \frac{\alpha^\alpha \nu_i^{\alpha-1} \exp(-\alpha\nu_i)}{\Gamma(\alpha)}, \end{aligned}$$

where $\lambda_i = \exp(\beta_1' \mathbf{z}_{1i})$, $\psi_i = \exp(\beta_2' \mathbf{z}_{2i})$ and $\mathbf{z}_{1i}$, $\mathbf{z}_{2i}$ are vectors of covariates, not necessarily distinct.
The joint model shows superiority over standard survival methods, the inclusion of additional information in the joint model resulted in an increase in power, which consequently meant that statistically significant covariate effects, not recognised by the standard survival distributions, could be affirmed. There are interesting characteristics within the data however, not present in the model, that were also highlighted in analysis. This paper will consider subsequent modifications to the joint model to accommodate these properties. It is important to note that

812 of the 1425 individuals included in the exploratory analysis presented only a single seizure pre-randomisation. The period of time from this single seizure to randomisation, for these individuals, ranged from the same day to 464 days, with the median number of days being 27. For the majority of those individuals with only one seizure pre-randomisation, their associated period of time from first seizure to randomisation may be inaccurately small, possibly representing how long it took for them to arrange an appointment with their GP. This results in imprecise estimates of their associated underlying seizure rates and an ensuing overestimation of the seizure rate reductions. Following discussions with clinicians, we subsequently made adjustments to the values of u_i in the data set so that $u_i \geq 182$. A sensitivity analysis was carried out to accompany these adjustments.

3 Extended and Cure Rate Models

3.1 Time Varying Post-Randomisation Rates

The results have suggested that the iid assumption for Y_{1i} and Y_{2i} may not be accurate. Instead we consider the following adjustment to the joint density for the post-randomisation survival times:

$$f_{Y_1,Y_2|\nu}(y_{1i}, y_{2i} \mid \nu_i; \lambda_i, \psi_i) \quad = \quad (\lambda_i \psi_{1i} \nu_i)^2 \psi_{2i} \exp(-\lambda_i \psi_{1i} \nu_i (y_{1i} + \psi_{2i} y_{2i})),$$

where $\lambda_i = \exp(\beta_1' \mathbf{z}_{1i})$, $\psi_{1i} = \exp\{\beta_2' \mathbf{z}_{2i}\}$, $\psi_{2i} = \exp(\beta_3' \mathbf{z}_{3i})$ and $\mathbf{z}_{1i}$, $\mathbf{z}_{2i}$, $\mathbf{z}_{3i}$ are vectors of covariates, not necessarily distinct.

Table 1 shows the subsequent estimated pre-randomisation seizure rates and the expected yearly seizure rates, stratified by seizure type. Neither sex or age were found to be significant in determining baseline seizure rates or post-randomisation seizure rate reductions, so we exclude these variables completely. Those with generalised seizures pre-randomisation generally have the highest seizure rate, but these individuals can generally expect to see the greatest reduction in seizure rates post-randomisation. Table 2 shows the maximum likelihood estimates for ψ_{1i} and ψ_{2i}. The values of ψ_{1i} represent the change in rate following randomisation, with ψ_{2i} representing the change in rate following first post-randomisation seizure. Treatment policy does not appear to be statistically significant for those individuals with a normal EEG. Additionally, those individuals having an abnormal EEG, but allocated to immediate treatment can expect to have a post-randomisation seizure rate in line with those presenting a normal EEG. For those with an abnormal EEG immediate treatment is favoured for all groups except partial, where no statistically significant difference between treatment policies is observed. Following a first seizure post-randomisation we see that in general seizure rates increase. We now see no significant differences in EEG outcomes and treatment policies.

3.2 Cure Rate Model

On average, around 50% of people do not experience seizure recurrence after a single, untreated epileptic seizure and recall that over half of the 1425 individuals

Seizure Type	$\widehat{\lambda}_i$ (95% C.I.)	Expected yearly rate
Tonic-Clonic	0.0054 (0.005,0.006)	2
2° Tonic-Clonic	0.008 (0.007,0.009)	3
Generalised	0.055 (0.044,0.070)	20
Partial	0.016 (0.013,0.019)	6
Other	0.022 (0.016,0.030)	8

TABLE 1. The expected pre-randomisation seizure rate per unit time.

Seizure Type	$\widehat{\psi}_{1i}$ (95% C.I.) first seizure			
	Abnormal EEG			
	Immediate		Deferred	
Tonic-Clonic	0.093	(0.07,0.12)	0.207	(0.16,0.26)
2° Tonic-Clonic	0.123	(0.09,0.16)	0.294	(0.22,0.39)
Generalised	0.021	(0.01,0.04)	0.091	(0.05,0.17)
Partial	0.056	(0.03,0.10)	0.069	(0.05,0.10)
Other	0.091	(0.03,0.27)	0.313	(0.11,0.86)
	Normal EEG			
	Immediate		Deferred	
Tonic-Clonic	0.085	(0.07,0.11)	0.111	(0.09,0.14)
2° Tonic-Clonic	0.055	(0.04,0.07)	0.077	(0.06,0.10)
Generalised	0.028	(0.01,0.06)	0.070	(0.03,0.15)
Partial	0.129	(0.067,0.24)	0.092	(0.05,0.18)
Other	0.048	(0.02,0.12)	0.095	(0.04,0.26)
	$\widehat{\psi}_{2i}$ (95% C.I.) second seizure			
	Abnormal EEG			
	Immediate		Deferred	
Tonic-Clonic	5.260	(3.53,7.83)	0.950	(0.67,1.35)
2° Tonic-Clonic	1.471	(0.96,2.25)	0.833	(0.54,1.28)
Generalised	2.968	(1.10,8.00)	3.051	(1.34,6.94)
Partial	2.033	(0.91,4.57)	1.845	(0.85,3.99)
Other	0.815	(0.06,12.01)	0.895	(0.17,4.82)
	Normal EEG			
	Immediate		Deferred	
Tonic-Clonic	3.658	(2.56,5.23)	1.801	(1.28,2.54)
2° Tonic-Clonic	4.095	(2.56,6.55)	6.324	(4.08,9.81)
Generalised	2.573	(0.66,10.09)	7.216	(1.77,29.49)
Partial	1.322	(0.52,3.36)	3.272	(1.29,8.31)
Other	0.476	(0.08,2.68)	1.426	(0.28,7.34)

TABLE 2. The expected change in seizure rates following randomisation and first post-randomisation seizure.

for which exploratory analysis was carried out presented only a single seizure pre-randomisation. It is therefore not unreasonable to suspect that there may be a substantial proportion of the sample likely to be 'immune' from future seizures post-randomisation. If survival data does indeed have a proportion that are immune to the event of interest, a model that ignores this may give misleading results. More specifically, ignoring any potential cure fraction could result in underestimates of the post-randomisation seizure rates, thus contributing to the magnitude of seizure rate reductions that have been observed.

We have adjusted our original model to allow for the inclusion of a cure fraction. We first consider a model that jointly models the pre-randomisation seizure counts and post-randomisation time to first seizure only, which has the following density and survivor function:

$$\begin{aligned} f_{Y_1|\nu}(y_{1i} \mid \nu_i; \lambda_i, \psi_i) &= p\lambda_i\psi_i\nu_i \exp(-\lambda_i\psi_i\nu_i y_{1i}), \\ S_{Y_1|\nu}(y_{1i} \mid \nu_i; \lambda_i, \psi_i) &= 1 - p + p\exp(-\lambda_i\psi_i\nu_i y_{1i}), \end{aligned}$$

where $\lambda_i = \exp(\beta_1' \mathbf{z}_{1i})$, $\psi_i = \exp(\beta_2' \mathbf{z}_{2i})$ and $\mathbf{z}_{1i}$, $\mathbf{z}_{2i}$ are vectors of covariates, not necessarily distinct. The term p represents the susceptible proportion in the population, so that $1 - p$ is the cure fraction.

Table 3 shows how the post-randomisation seizure rate reductions are affected by the inclusion of cure rates in the model. The maximum likelihood estimate for p was 0.586 (standard error 0.02), with a highly statistically significant likelihood-ratio test statistic of 442. This value suggests that there is a substantial proportion of the population 'immune' from seizures post-randomisation.

If we look at Table 3 we can see that after incorporating a cure fraction into the model, seizure type is no longer significant in the magnitude of seizure rate reductions, apart from those with generalised seizures, with an abnormal EEG and allocated to deferred treatment. We can also see that those individuals with an abnormal EEG, randomised to the deferred treatment group can in general expect to experience seizures around 70% as often as pre-randomisation; additionally those with 2° Tonic-Clonic, Partial or Other seizures, the estimated ψ_i is not statistically significantly different from 1, which would correspond to no change in seizure rate. It is interesting to note that those individuals allocated to deferred treatment, but with a normal EEG still seem to have a substantial reduction in their seizure rates post-randomisation.

4 Conclusions

Our initial joint model of pre-randomisation and post-randomisation seizure rates provides an improvement over standard survival models. Despite an observed improvement, the joint model does not incorporate other characteristics within the data evident in exploratory analyses. Ignoring these characteristics can give misleading results and whilst the current model is useful for presenting comparisons between seizure type groups and treatment groups, the results obtained from the simple joint model do not allow us to formulate absolute risks of future seizures for patients with early epilepsy.

In order to enhance understanding of the problem and aid clinicians and patients in their treatment decisions, we applied the extensions discussed and observed

Seizure Type	$\widehat{\psi}_{1i}$ (95% C.I.)			
	Abnormal EEG			
	Immediate		Deferred	
Tonic-Clonic	0.248	(0.17,0.37)	0.662	(0.50,0.88)
2° Tonic-Clonic	0.230	(0.16,0.32)	0.824	(0.60,1.13)
Generalised	0.140	(0.07,0.30)	0.144	(0.08,0.25)
Partial	0.477	(0.26,0.86)	0.682	(0.39,1.19)
Other	0.154	(0.02,0.96)	0.670	(0.23,1.99)
	Normal EEG			
	Immediate		Deferred	
Tonic-Clonic	0.417	(0.30,0.58)	0.590	(0.43,0.80)
2° Tonic-Clonic	0.251	(0.16,0.41)	0.476	(0.32,0.70)
Generalised	0.381	(0.15,0.99)	0.206	(0.07,0.62)
Partial	0.471	(0.24,0.94)	0.356	(0.18,0.72)
Other	0.230	(0.06,0.83)	0.531	(0.19,1.52)

TABLE 3. The expected change in seizure rate post-randomisation, under cure rate model.

improvements on our initial model. There was sufficient evidence to support the inclusion of varying seizure rates post-randomisation and evidence to suggest that there may a substantial cure fraction present in the population. Further extensions that now need to be considered are extending the cure rate model to include both post-randomisation survival times and allowing cure rates to depend on individuals' covariates. We would then like to combine the two extensions discussed in this paper so that we build a model that allows the post-randomisation seizure rates to vary and incorporates the appropriate cure fractions.

Keywords: Epilepsy; Recurrent events; Survival analysis.

References

Cowling, B. J. and Hutton, J. L. and Shaw, J. E. H. (2006). Joint modelling of event counts and survival times. *Journal of the Royal Statistical Society, Series C*, **55**, 31-39.

Marson, A. and Jacoby, A. and Johnson, A. and Kim, L. and Gamble, C. Chadwick, D. (2005). Immediate versus deferred antiepileptic drug treatment for early epilepsy and single seizures: a randomised control trial. *The Lancet*, **365**, 2007-2013.

Rogers, J. K. and Hutton, J. L. and Hemming, K. (2009). Joint Modelling of Event Counts and Survival Times. CRiSM Working Paper, 44 *http://www2.warwick.ac.uk/fac/sci/statistics/crism/research/2009/paper09-44*. University of Warwick.

Estimation of multiple correlated effects on a disease outcome for multilevel data

Giulia Roli[1]

[1] Department of statistical sciences, University of Bologna, via Belle Arti, 41, 40126, Bologna. E-mail: g.roli@unibo.it

Keywords: disease outcome; hierarchical Bayesian models; correlated exposures; multilevel data.

1 Introduction and aims

When a case-control study aims to investigate the exposures which can be the cause of the occurrence of a disease, epidemiologists often deal with some complications that need to be somehow controlled during the analysis. In this paper, we consider two kinds of such complications. The first one concerns the multilevel structure of the data, that is subjects nested into higher level units involving their own variability. As a consequence, the within-cluster dependence among the observations is neither accidental nor ignorable. Moreover, the risks of drawing wrong conclusions are high if the clustering of the data is disregarded (Raudenbush and Bryk (2002)).
The joint analysis of multiple exposures gives rise to the second complication. Indeed, many epidemiologic studies involve a set of potential effects to be compared and, as a result, face problems of multiple inference (Thomas et al. (1985)). When a conventional analysis is carried out, these problems are revealed by failures in the convergence of the estimation process or by implausible large and unstable estimates, especially when the samples are small and sparse (Greenland (1992), Witte et al. (1994)). The main reason is that these effects are often correlated. Therefore, we need to take into account for a covariance structure among them to reduce the random errors in the estimates.
We propose a hierarchical Bayesian model to tackle both these complications. Additional information and previous knowledge are exploited to specify reasonable prior assumptions on the crucial parameters. This approach allows to address the problems of sparse data and study bias, as well as multiple comparisons and subgroup analysis.
The model is developed to be applied to a study aiming to investigate the association of dietary exposures with the occurrence of colon-rectum cancer. A multilevel setting is involved, as individuals have been enrolled from different countries and centers of Europe. Additional data on the nutrient

compositions of each dietary item are arranged to model the correlation among the exposures and to improve the estimates.

2 Data and methods

We consider a sample of 24,376 individuals nested in 27 European centers of recruitment drawn from the European Prospective Investigation into Cancer and Nutrition (EPIC) study. Subjects who developed a colon-rectum cancer after the enrollment are included in the analysis. Then, a number of controls are randomly selected to be equal to 5% of the whole set of control units, separately by center.

Dietary information for each subject are collected during the enrollment. A list of 30 food groups are considered, where the corresponding individual intakes are expressed in grams-per-day (gm/d). Some potential confounders (e.g., sex, gender, smoking status, etc.) are further included into the analysis to control for their effects.

Additional data on the nutrient compositions of each dietary exposure are available. In detail, these concern the amounts of constituents for one gram of each food. These data are arranged in matrices where the generic k-th row refers to the amounts of food constituents for the k-th dietary exposure. Such matrices are usually named tables of nutrient composition and may vary between countries and centers. As a result, they can be generally regarded as center-specific information which can further contribute to model the variability among the centers. According to the dietary items involved into the analysis, we select a list including the most considerable nutrients.

The hierarchical Bayesian model we propose involves 3 levels. The first one describes the conditional likelihood model for the observable disease indicator Y_{ij} for the individual i in the center j in the form of a conventional center-specific logistic regression:

$$logit[E(y_{ij}|x_{kij}, w_{pij}, \alpha_j, \beta_{kj}, \gamma_p)] = \alpha_j + \sum_{k=1}^{K} \beta_{kj} x_{kij} + \sum_{p=1}^{P} \gamma_p w_{pij} \quad (1)$$

where the intercepts α_j and the dietary effects β_{kj} are assumed to be random across the centers; x_{kij} is the individual dietary intake for the k-th food group in the center j; γ_p is the fixed effect of the p-th potential confounder; and w_{pij} represents the individual information about the p-th confounder in the j-th center.

At level 2, two sets of linear regressions are considered. We assume an empty model for the intercepts which splits the random parameter into a common effect, ψ_0, and a residual term, u_j, yielding the differences among the centers:

$$\alpha_j = \psi_0 + u_j \quad (2)$$

where the u_j are assumed to be independent and normally distributed with null means and common variances ϕ^2.
The data on constituents for each food are used to develop the level-2 model for the dietary coefficients. In detail, these are regressed on the nutrient covariates z_{qkj} as follows:

$$\beta_{kj} = \pi_0 + \sum_{q=1}^{Q} \pi_q z_{qkj} + \delta_{kj} \tag{3}$$

where we assume that the effects of the food exposures on the colon-rectum cancer are partially mediated by the effects of nutrients π_q. The residuals δ_{kj} are independent normal random variables with zero means and variances τ^2. We impose an independence structure for the parameters δ_{kj} which implies that any prior correlations among dietary effects are entirely explained by known differences in their constituents. Moreover, once the correlations of the exposures are modeled, the variability among groups is supposed to be negligible or null.
Level 3 completes the hierarchical structure of the parameters in a Bayesian framework by assigning hyper-prior distributions on the parameters ψ_0, γ_p, ϕ^2, π_q and τ^2. We impose an informative distribution only for the parameter τ^2 according to plausible range of variation for log normal random effects. Indeed, the role of this parameter is of crucial importance as it represents the uncertainty about the residual δ_{kj} and, from a Bayesian perspective, the estimate of τ^2 indexes different opinions about δ_{kj}. Moreover, previous works (Witte et al. (1994)) have shown that pre-specifying the value of τ^2 or its distribution yields better estimates when the sample size and the ratio of subjects to parameters are not large. In our case, we believe that a 2-fold variation between the Odds Ratios (OR) of dietary effects for the upper and lower 5% of units is reasonable and that a 4-fold variation between the upper and the lower 5% of units is very unlikely (say, less than a 1% chance). Both these restrictions allow us to specify a proper hyper-prior distribution (Gelman et al. (2003)) for the precision τ^{-2}, i.e., a Gamma distribution with parameters 5 and 0.22.

3 Results

In order to measure the improvement in the estimates of dietary effects, we compare the results from this hierarchical Bayesian model with those obtained by carrying out several conventional logistic regressions separately by center of enrollment j.
Some notable results are showed in Table 1, where the ORs and their 95% confidence intervals (CI) are calculated according to food-specific values of unit increase which are the sample standard deviations.
The results from the conventional disease model are notably affected by problems of sparse data which preclude the full estimation of each dietary

TABLE 1. Main results.

Dietary	Center	Conventional logistic model OR (CI)	Hierarchical Bayesian model OR (CI)
Milk	Navarra	-	0.900 (0.779-1.033)
Legumes	Granada	-	0.943 (0.826-1.075)
Cabbages	Turin	5.503 (0.089-340.06)	0.969 (0.836-1.128)
Processed meat	SC of France	1.901 (0.777-4.654)	1.076 (0.934-1.241)
Milk	NW of Norway	2.275 (0.620-8.340)	0.964 (0.833-1.119)
Milk	Malmo	0.999 (0.853-1.171)	0.962 (0.870-1.063)
Legumes	San Sebastian	0.938 (0.756-1.160)	0.974 (0.888-1.061)
Fish	Copenhagen	0.848 (0.693-1.037)	0.889 (0.799-0.989)
Processed meat	Malmo	1.137 (0.989-1.306)	1.110 (1.011-1.213)
Fruits	Aarhus	0.660 (0.492-0.885)	0.883 (0.779-0.995)

effect on the occurrence of colon-rectum cancer. In some cases, the ML estimation fails to converge because the predictors are highly correlated. Even when the convergence is achieved, a great number of estimates result with large and unstable absolute values, suggesting implausible strong associations according to the relevant diet and colon-rectum cancer literature.

When the hierarchical Bayesian model is fitted, formerly extreme and unstable estimates become more reasonable and less biased, even when the results on the same exposure are compared across different centers. This improvement is mainly due to the shared food information on nutrients also across different centers. As a result, the dietary estimates are pulled toward each other when they have similar compositions. This shrinkage is expected to occur especially for the same exposures evaluated in different centers as their levels of nutrients are more likely to be similar. Indeed, previous evidence (Roli (2006)) showed that the substantive improvements in the estimation of dietary effects are gained when a single multilevel analysis is carried out, while the inclusion of nutrient information alone for separate conventional regressions does not yield as good results. On the other hand, moderate and stable estimates remain much more the same, apart from great gains in terms of standard errors.

The shrinkage of the estimates can be evaluated in practice by plotting the results from the conventional regressions and from the hierarchical Bayesian method, simultaneously (Figure 1). Indeed, for the former we can observe a great variability with peaks of extremely high and extremely low numbers. Conversely, the estimates from our model are closer to each other (i.e., to the prior means based on the nutrients) and are controlled for variations due to random occurrences in small samples.

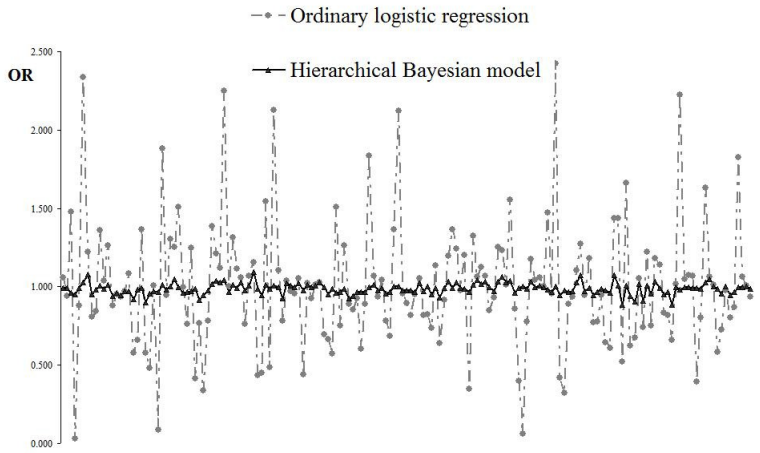


FIGURE 1. Estimated ORs

4 Conclusions

Statistical theory, several simulation studies and a large number of applications all support the use of hierarchical modeling as a powerful method which allows to yield strong gains in the accuracy of predictions and effect estimates. The improvement is mainly due to the use of prior data arranged in an additional model. As a result, the ordinary estimates from the conventional level-1 model are pulled or 'shrunk' toward each other when they have similar levels of prior data.

In multiple regression analysis, the hierarchical framework can further provide an alternative to conventional variable selection techniques. These procedures begin with a maximal model including all the terms (such as in backward elimination) or a minimal model that has only the essential regressors, i.e. the confounders, (such as in forward and stepwise selection) and proceed with a model reduction based on some significance criteria to search for a final model. The hierarchical approach states the maximal model as the level-1 regression. Then, it specifies a level-2 model, where the corresponding values of the residual variances mark the degree of compromise between the extremes of putting each variable completely in or completely out of the model. Therefore, when these level-2 variances are null, then a minimal model holds; conversely, if they are large, the final model tends to be the maximal one. Moreover, the hierarchical approach does not make a definitively "all-or-nothing" choice for each term, but allows to retain all the variables in the analysis in order to be further evaluated whenever additional information would be available.

The advantages related to the use of hierarchical methods under a Bayesian setting are highlighted by the results of the empirical illustration, where for a multicentric study the ordinary ML estimates of multiple dietary effects are improved, for each center separately, by a hierarchy of models merging and exploiting all the prior knowledge about the problem at hand. The improvement is expressed in terms of more plausible estimates of dietary effects and lower mean-squared errors than traditional data summaries, thanks to a two-fold shrinkage action due to the similar nutrient compositions of dietary items between and within the centers.
The hierarchical Bayesian model we propose can be further applied in many other epidemiologic contexts. For instance, in occupational studies, where more levels of information can be merged; or to perform polytomous logistic regressions of different causes of death on a set of exposures; or in disease mapping and spatial analysis, where the variations due to random occurrences need to be controlled by exploiting the spatial proximity and the consequent interaction of the geographical areas.
In all these examples, the use of hierarchical Bayesian modeling can be easily extended or raised thanks to the substantial gains that it can yield and its internal flexibility as regards the prior assumptions. This paper is intended to encourage the use of Bayesian methods in epidemiology as a powerful statistical tool to address the problem of nested data and correlated effects.

References

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003). *Bayesian Data Analysis*. 2nd ed., Chapman and Hall/CRC.

Greenland, S. (1992). A semi-bayes approach to the analysis of correlated multiple associations, with an application to an occipational cancer-mortality study. *Statistics in medicine*, **11**, 219230.

Raudenbush, S., and Bryk, A. (2002). *Hierarchical Linear Models - Application and data analysis methods*. 2nd ed., Sage.

Roli (2006). Hierarchical logistic regression in a multicentric study of multiple dietary effects on a disease outcome: a fully Bayesian approach. PHD thesis.
(Available from www.stat.unibo.it/ScienzeStatistiche/Ricerca/Dottorati).

Thomas, D., Semiatycki, J., Dewar, R., Robins, J., Goldberg, M., and Armstrong, B. (1985). The problem of multiple inference in studies designed to generate hypotheses. *American journal of epidemiology*, **122**, 10801095.

Witte, J., Greenland, S., Haile, R., and Bird, C. (1994). Hierarchical regression analysis applied to a study of multiple dietary exposures and breast cancer. *Epidemiology*, **5**, 612621.

Heteroscedastic nonlinear elliptical models for correlated data

Cibele M. Russo[1], Gilberto A. Paula[1], Francisco José A. Cysneiros[2], Reiko Aoki[3]

[1] Departamento de Estatística, Instituto de Matemática e Estatística, Universidade de São Paulo, Caixa Postal 66281 (Ag. Cidade de São Paulo), CEP 05311-970, São Paulo, SP, Brazil, e-mail: cibele@ime.usp.br and giapaula@ime.usp.br

[2] Departamento de Estatística, Universidade Federal de Pernambuco, CEP 50749-540, Recife, PE, Brazil, e-mail: cysneiros@de.ufpe.br

[3] Departamento de Matemática Aplicada e Estatística, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Caixa Postal 668, CEP 13560-970, São Carlos, SP, Brazil, e-mail: reiko@icmc.usp.br

Abstract: In this work we propose heteroscedastic nonlinear elliptical models with random effects. Our aim is to extend the model proposed by Russo et al. (2009) supposing heteroscedastic structures for the scale matrix, since the dispersion of the data may vary in different levels of some covariate, for instance the time. To provide a better control on the sources of variation in the model, we also verify the presence of autocorrelation between measurements. Moreover, the elliptical distributions are assumed for the joint distribution of random effects and errors, which allow the attribution of different weights to the observations. As numerical illustration we consider the indomethacin concentration data set, previously analysed by Bocheng & Xuping (2001).

Keywords: nonlinear models; elliptical distributions; heteroscedastic models; autoregressive structure; random effects.

1 Introduction

When the dispersion of the data is not homogeneous over time, for instance, the assumption of homoscedasticity may be unappropriate. Particularly for longitudinal data as repeated measures or growth curves, heteroscedastic and/or autoregressive structures are an alternative to control different sources of variation in the data. In this work we also assume that the random effects and errors follow an elliptical multivariate distribution, which may provide more flexible fits since this class includes light- and heavy-tailed distributions such as Student-t, power exponential, logistic, normal, among others (see, for instance, Fang et al., 1990). Recently, Russo et al. (2009) proposed nonlinear elliptical models with mixed-effects for longitudinal data. In this work we propose an extension to these models assuming

heteroscedastic and/or autoregressive structures, depending on the necessity of the data. Recent works on heteroscedastic or autoregressive elliptical models are Paula et al. (2009) and Cysneiros et al. (2007).

2 Heteroscedastic models with autoregressive elliptical errors

Let $\mathbf{y}_i$ be an m_i-dimensional vector such that $\mathrm{E}(\mathbf{y}_i) = \boldsymbol{\mu}_i = \mathbf{f}(\mathbf{t}_i, \boldsymbol{\beta})$, for $i = 1, \ldots, n$. A possible mixed-effects model for $\mathbf{y}_i$ proposed by Russo et al. (2009) is given by

$$\mathbf{y}_i = \mathbf{f}(\mathbf{t}_i, \boldsymbol{\beta}) + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{1}$$

with $\mathbf{f}(\mathbf{t}_i, \boldsymbol{\beta}) = (f(t_{i1}, \boldsymbol{\beta}), \ldots, f(t_{im_i}, \boldsymbol{\beta}))^T$ being an m_i-dimensional nonlinear function of $\boldsymbol{\beta}$, $\mathbf{t}_i$ is a vector of explanatory variable values, $\mathbf{Z}_i$ is a matrix of known constants, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ a vector of unknown parameters, $\mathbf{b}_i = (b_{i1}, \ldots, b_{ir})^T$ a vector of unobserved random regression coefficients. In this paper we generalize this model by considering that

$$\begin{bmatrix} \mathbf{y}_i \\ \mathbf{b}_i \end{bmatrix} \sim \mathrm{El}_{m_i+r} \left\{ \begin{pmatrix} \mathbf{f}(\mathbf{t}_i, \boldsymbol{\beta}) \\ \mathbf{0} \end{pmatrix}; \begin{bmatrix} \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \sigma^2 \mathbf{Q}^{-1} \mathbf{M}(\delta, \mathbf{t}_i) \mathbf{Q}^{-T} & \mathbf{Z}_i \mathbf{D} \\ \mathbf{D} \mathbf{Z}_i^T & \mathbf{D} \end{bmatrix} \right\}, \tag{2}$$

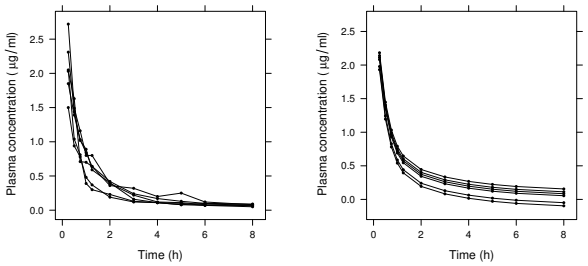
where the scale matrices $\boldsymbol{\Sigma}_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \sigma^2 \mathbf{Q}^{-1} \mathbf{M}(\delta, \mathbf{t}_i) \mathbf{Q}^{-T}$, $\mathbf{D}$, and $\mathbf{Z}_i \mathbf{D}$ are proportional to the variance-covariance matrices $\mathrm{Var}(\mathbf{y}_i)$, $\mathrm{Var}(\mathbf{b}_i)$ and $\mathrm{Cov}(\mathbf{y}_i, \mathbf{b}_i)$. The variance structure defined by $\sigma^2 \mathbf{Q}^{-1} \mathbf{M}(\delta, \mathbf{t}_i) \mathbf{Q}^{-T}$ was discussed in Lin & Wei (2003), where $\mathbf{M}_i = \mathbf{M}(\delta, \mathbf{t}_i)$ and $\mathbf{Q}$ introduces heteroscedasticity and autocorrelation, respectively.

Following Russo et al. (2009), we consider the marginal model, namely $\mathbf{y}_i \sim \mathrm{El}_{m_i}(\mathbf{f}(\mathbf{t}_i, \boldsymbol{\beta}); \boldsymbol{\Sigma}_i)$, which may be obtained without requiring numerical integration. To decide about the most appropriate model, information criteria as AIC and BIC may be used, and residual graphics may confirm the adequacy of the model. The maximum likelihood estimates of the parameters of interest may be easily obtained by using the Fisher scoring algorithm, which is similar to the iterative process presented in Russo et al. (2009), as well as the procedure for the estimation of random effects by using the empirical Bayes method.

3 Application

Bocheng & Xuping (2001) analysed a pharmacokinetics data set where the plasma concentration of indomethacin is nonlinearly related to the time after the injection, in which 11 measurements of indomethacin concentration in each of 6 volunteers with time varying from 15 minutes to 8 hours after the bolus intravenous injection with the same dose in each of the 6 volunteers were taken. To relate the plasma concentration y to

the time t, they considered the nonlinear model $\mathbf{y}_i = \mathbf{f}(\boldsymbol{\beta}, \mathbf{t}_i) + \mathbf{u}_i + \boldsymbol{\epsilon}_i$, in which $\mathbf{y}_i$ is the observation vector of indomethacin concentrations for the ith subject, $\mathbf{f}(\boldsymbol{\beta}, \mathbf{t}_i)$ is a 11×1 vector with each element given by $f(\boldsymbol{\beta}, \mathbf{t}_{ij}) = e^{\beta_1} \exp(-e^{\beta_2} t_{ij}) + e^{\beta_3} \exp(-e^{\beta_4} t_{ij})$, $\boldsymbol{\epsilon}_i$ and $\mathbf{u}_i$ are respectively the vectors of errors and random effects of the model. Bocheng & Xuping (2001) assume that $\boldsymbol{\epsilon}_i \overset{i.i.d}{\sim} N(0, \sigma^2\mathbf{I})$, $\mathbf{y}_i|\mathbf{u}_i \overset{i.i.d}{\sim} N(\mathbf{f}(\boldsymbol{\beta}, \mathbf{t}_i) + \mathbf{u}_i, \sigma^2\mathbf{I})$ and $\mathbf{u}_i \overset{i.i.d}{\sim} N(0, \sigma^2\boldsymbol{\Sigma})$.



(a) (b)

FIGURE 1. Indomethacin concentration against time (a) and estimated profiles under the model fitted by Bocheng & Xuping (2001)

The data are presented in Figure 1 (a), where it is possible to observe that the dispersion of the measurements of the plasma indomethacin concentration is larger in the beginning of the study (right after the injection) than after the time has passed, specially in the end of the study. The fitted profiles presented in Figure 1 (b) show that the model considered by Bocheng & Xuping (2001) does not seem to fit the data well, as it does not take into account the differences in the variability of the response variable for different times and the graphic of the standardized residuals against the independent variables (omitted here) also show a pattern of heteroscedasticity. Several models of the type of (1-2) were fitted, including heteroscedastic and/or autoregressive structures and different forms for $\mathbf{Z}_i$. The most appropriate model for the indomethacin data, according to AIC and BIC, was the one which is called the first-order approximation heteroscedastic autoregressive models (FOAHARM$_{\beta_1\beta_2}$), in which $\mathbf{Z}_i = [\partial\mathbf{f}(\mathbf{t}_i, \boldsymbol{\beta})/\partial\beta_1, \partial\mathbf{f}(\mathbf{t}_i, \boldsymbol{\beta})/\partial\beta_2]$ evaluated in the least squares estimate $\widetilde{\boldsymbol{\beta}}$. For this data set we considered $\mathbf{M}_i = \text{diag}[\exp(\delta\mathbf{t}_i)]$ and $\mathbf{Q}$ the standard matrix to introduce first-order autoregressive structure. More details of the fitted models are presented in Table 3.

The fitted profiles and residuals graphics (omitted here) from the heteroscedastic elliptical models supposing heavy-tailed distributions seem to

TABLE 1. Parameter estimates under normal, Student-t and power exponential FOAHARM$_{\beta_1\beta_2}$'s for indomethacin data

	Normal		Student-t		Power exponential	
	Estimate	(SE)	Estimate	(SE)	Estimate	(SE)
β_1	1.0198	(0.0860)	0.9895	(0.0866)	0.9754	(0.0878)
β_2	0.6195	(0.0794)	0.5895	(0.0758)	0.5937	(0.0755)
β_3	-1.2274	(0.2353)	-1.3828	(0.2614)	-1.3797	(0.2643)
β_4	-1.7208	(0.1582)	-1.8127	(0.186)	-1.7982	(0.1843)
σ^2	0.0323	(0.0103)	0.0256	(0.0098)	0.0001	(0.0000*)
τ_1	0.0279	(0.0245)	0.0272	(0.0242)	0.0001	(0.0001)
τ_2	0.0705	(0.0495)	0.0737	(0.054)	0.0004	(0.0003)
δ	-0.7295	(0.0808)	-0.7279	(0.0847)	-0.7227	(0.0851)
ρ	0.1466	(0.0038)	0.1638	(0.0031)	0.1508	(0.0000*)

* greater than zero

provide more appropriate fits to the data set. The presented results are valid for other autoregressive structures and/or other heteroscedasticity functions.

Acknowledgments: The authors are grateful to FAPESP, FACEPE and CNPq, Brazil, which supported this research.

References

Bocheng, W. and Xuping, Z. (2001). Influence analysis in nonlinear models with random effects. *Applied Mathematics. A Journal of Chinese Universities. Series B*, **16**, 35–44.

Cysneiros, F. J. A., Paula, G.A. and Galea, M. (2001). Heteroscedastic symmetrical linear models, *Statistics and Probability Letters*, **77**, 1084–1090.

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

Lin, J.G. and Wei, B.C. (2003). Testing for heteroscedasticity in nonlinear regression models. Communications in Statistics - Theory and Methods, 32, 171–192.

Paula, G.A. and Medeiros, M. and Vilca-Labra, F. E. (2009). Influence diagnostics for linear models with first-order autoregressive elliptical errors. *Statistics and Probability Letters*, **79**, 339–346.

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

Haulm senescence in potatoes and semi-parametric survival models

Sabine K. Schnabel[1,3], Paul H.C. Eilers[1,2],Paula Hurtado López[1,4,5], Richard G.F. Visser[3,4], Fred A. van Eeuwijk[1,3]

[1] Biometris, Wageningen UR, Postbus 100, 6700 AC Wageningen, The Netherlands; sabine.schnabel@wur.nl (communicating author)
[2] Erasmus MC, Department of Biostatistics, Postbus 2040, 3000 CA Rotterdam, The Netherlands
[3] Center for Biosystems Genomics, Postbus 98, 6700 AB Wageningen, The Netherlands
[4] Wageningen UR, Plant Breeding, Postbus 386, 6700 AJ Wageningen, The Netherlands
[5] C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC), Wageningen UR, Droevendaalsesteeg 4, 6708 PB Wageningen, The Netherlands

Abstract: Haulm senescence describes the decay of potato plants on a discrete visual scale. A smooth semi-parametric hazard model is developed and applied to data from an agricultural experiment. Characteristics of the estimated hazard and survival function can be used to detect quantitative trait loci on chromosomes that can be associated with the senescence process.

Keywords: Hazard, P-splines, Poisson distribution, senescence QTL analysis

1 Introduction

An important phase in the life of a (commercial) potato plant is the decay of the haulm, the part above the ground. Only after complete decay potatoes will be harvested. In agricultural experiments, the state of the haulm, called its senescence, is monitored regularly and visually graded on a discrete scale. Breeders are interested in characterizing haulm senescence and in finding genes that influence it.
Traditionally senescence data have been modeled by parametric curves (Malosetti et al., 2006). In addition we have been working on semi-parametric alternatives (Hurtado at al., 2010). They increase the flexibility of the statistical model, but treat the data as independent observations of a time series. However, senescence is an irreversible process, so this model is lacking a strong biological basis. Here we introduce a semi-parametric survival model that is more realistic.

To improve the yield of potato plant breeders are interested in identifying the genomic regions underlying yield and yield-related traits. Such a region is called a quantitative trait locus (QTL). Plants that senesce more slowly retain higher photosynthesic capacity and thereby may reach higher yields. Statistical descriptions of the senescence process become useful for plant breeding purposes when they deliver process characterizations that are clearly and consistently different between potato plants with different genetic constitutions (genotypes). Identification of the genomic regions underlying senescence differences, so called QTL mapping, can be understood as the comparison between groups of genotypes with different DNA composition at specified positions in the genome for particular response traits such as yield and senescence characterizations. The statistical modeling and characterization of the senescence process is more successful when more and stronger QTLs are detected. For an example data set we show that our semi-parametric survival analysis approach was able to produce senescence characterizations for which clear QTLs turned up.

2 Data description

Our senescence data were obtained from a Finnish field experiment in 2004 with around 200 genotypes of a diploid potato population (Zaban et al., 2006). The senescence process of the haulms was recorded at different time points during the growing season. The status of the haulm y_i was scored on a scale from "green plant" ($y_i = 0$), "upper leaves with first signs of yellowing" ($y_i = 1$) etc. to "dead plant" ($y_i = 7$) (Celis-Gamboa et al., 2003).

To translate the senescence data to a survival model, we introduce an urn with seven marbles. At each step on the senescence scale a marble is removed. If monitoring had been done daily and long enough, we would know the times at which each change of state (removing of a marble) occurred. Then we would have had discrete survival data. In reality observations were made intermittently, approximately every 5 days (counted in days after planting: DAP). So the data are interval censored. We presently handle this complication by simply assuming the observed changes (zero, one, or more than one marble removed) to have occurred uniformly distributed over the respective time interval. We will return to this in the Discussion.

For each genotype three replicated plants were monitored. To simplify the analysis, their results were combined into one hypothetical urn with three times seven marbles.

The genotypes in the field experiment show a lot of variation in their behavior in terms of senescence as well as in other observed traits of the plants. Some genotypes develop early and therefore reach the state of "dead plant" (equivalent to zero marbles in the urn), while others develop late and might not reach the last state of the senescence scale. This diversity can also be seen in the final results.

3 Theory and Application

The time axis is divided into narrow intervals. In interval j we have the number at risk r_j and the number of events d_j. We model the logarithm of the hazard:

$$\log h_j = \sum_k b_{jk}\alpha_k, \tag{1}$$

where $B = [b_{jk}]$ is the matrix of $B-$spline basis functions and α the coefficient vector. The log-likelihood is

$$\log L = \sum_j d_j \log \mu_j - \sum_j \mu_j \tag{2}$$

with $\mu_j = \mathbf{E}(d_j) = r_j h_j$ the expected value of the number of events in interval j. $P-$splines include a penalty $\lambda||D_3\alpha||^2$ to get a smooth curve, where D_3 is a matrix that forms third order differences and λ a parameter to tune smoothness. After linearization we find that we have to solve the following system repeatedly:

$$(B^T\tilde{M}B + \lambda D_3^T D_3)\alpha = B^T(d - \tilde{\mu} + \tilde{M}B\tilde{\alpha}). \tag{3}$$

where $\tilde{M} = \mathrm{diag}(\tilde{\mu})$. A more detailed description can be found in (Eilers, 1998).

In our application to haulm senescence of potato we estimate an individual hazard curve for each genotype in the population. Five examples of estimated smooth hazard curves along with their respective fitted and empirical survival curves are shown in Figure 1. As mentioned above we can see quite a range of different shapes for the hazard curves as well as for the survival curves. While the first two genotypes (CE140 and CE691) are commonly classified as intermediate, the other genotypes are developing early. However, we see that in this group the shape of hazard curves varies substantially. While CE102 and CE155 show a unimodal shape, CE685 tends to a bimodal hazard. In this case the senescence process seems to level off after an initial period of aging. The hazard increases then again towards the end of the observation period.

3.1 Analysis of quantitative trait loci

In order to identify possible QTLs influencing haulm senescence we can use the fitted curves. To characterize the curves we determine the mode of the hazard and the time point when it occurs as well as the time point when 1/5 of the senescence process is over (indicated by the horizontal lines in the survival curves of Figure 1). With these characteristics we performed a non-parametric QTL analysis using the rank sum test of Kruskal-Wallis

available in MapQTL 6 (van Ooijen, 2009). QTL mapping was performed separately on the maternal and paternal maps (C and E respectively) and the criterion for detecting QTLs was set by a significance level of $p \leq 0.005$. We detected a major QTL on chromosome 5 related with the mode of the hazard, a QTL on chromosome 4 of the C parent related with the time point when 1/5 of the senescence process occurred as well as minor QTLs on other chromosomes. These findings are in line with previous research (Hurtado et al., 2010). The QTL on chromosome 4 related with the time point when 1/5 of the senescence process elapsed leads to an interpretation of that time point as the onset of senescence.

4 Discussion

To our knowledge, survival models have not been used for haulm senescence. We believe they offer a more realistic basis for the analysis of development traits over time than a curve fitting approach whether parametric (Malosetti et al., 2006) or semi-parametric (Hurtado at al., 2010). On the other hand we are aware that we only have scratched the surface of this field. Here we shortly discuss in which directions our research will be extended.
In this analysis we treated interval censoring in a very simplified manner and assumed a uniform distribution of events inside a time interval. In principle it is straightforward to write down a model in which a smooth hazard *and* a smoothly changing risk set (derived from that hazard) define the likelihood.
All marbles in the urn were considered exchangeable. In reality we have a multi-state model with a chain of states. The relative hazards of the changes of states most probably are not equal. By combining data from different genotypes in the same experiment, or from the same genotype in different experiments, or both, it might be possible to estimate an inflation or deflation factor for each state that is modulated by the overall smooth hazard.
Haulm senescence is strongly influenced by temperature. A real challenge will be to develop models in which genotype-specific parameters and observed temperature are combined. This will involve large-scale mixed model technology to combine data from different genotypes and different environments. Generalized additive models for the log-hazard, with additional shrinkage, might be a promising starting point.
Furthermore we also want to explore transformations of the time axis. The conventional choice of "days after planting" as time scale does not take into account temperature or photo period. We are currently experimenting with time measures including these factors such as thermal time and extensions to it. Thermal time is a measure for daily accumulations of heat taking into account the minimum and maximum temperature for growth of the particular species. This cumulative measure for heat can be further

extended to include the daylight length which is another important factor in the development of the plant.
We will also need good summaries of the results from our semi-parametric survival model to serve the needs of potato breeders well. In the present model the height, the position and the width of the hazard curve are candidates, but in more complex model the choice of characterizations may be less obvious.

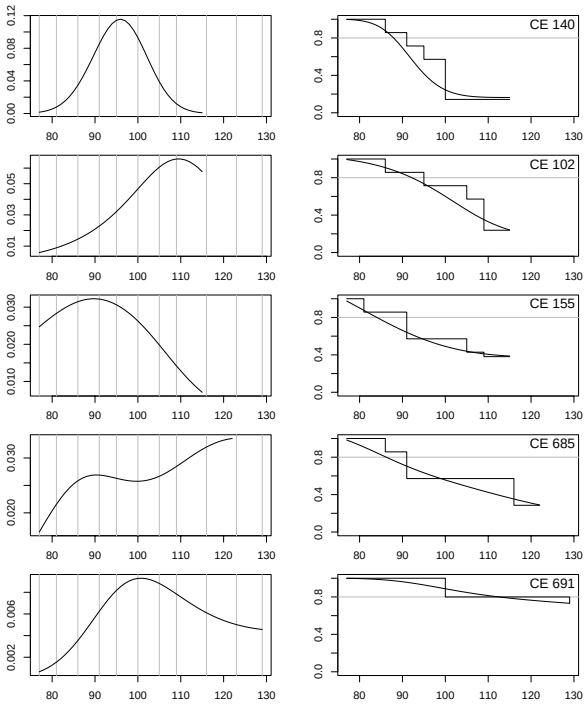


FIGURE 1. Left side: Estimated smooth hazard curves for five selected genotypes over days after planting (DAP). Vertical lines indicate the observation time points. Right side: Empirical (thin line) and fitted survival (thick line) for five selected genotypes over DAP. Grey line indicates 80 % survival.

References

Celis-Gamboa, C., Struik, P.C., Jacobsen, E., and Visser, R.G.F. (2003). Temporal dynamics of tuber formation and related processes in a

crossing population of potato (Solanum tuberosum). *Annals of Applied Biology*, **143**, 175-186.

Eilers, P.H.C. (1998). Hazard smoothing with B–splines. In: Statistical Modeling. Proceedings of the 13th International Workshop on Statistical Modelling. 200-207, New Orleans, USA.

Hurtado, P., Schnabel, S., Zaban, A., Veteläinen, M., Virtainen, E. , Eilers, P., van Eeuwijk, F., Visser, R., and Maliepaard, C. (2010). Dynamics of senescence-related QTL in potato. *Under review*.

Malosetti, M., Visser, R.G.F., Celis-Gamboa, C., and van Eeuwijk, F.A. (2006). QTL methodology for response curves on the basis of non-linear mixed models, with an illustration to senescence in potato. *Theoretical Applied Genetics*, **113**, 288-300.

van Ooijen, J.W. (2009). MapQTL 6, Software for the mapping of quantitative trait loci in experimental populations of diploid species. Kyazma B.V., Wageningen, Netherlands.

Zaban, A., Veteläinen, M., Celis-Gamboa, C.B., van Berloo, R., Häggman, H., Visser, R.G.F. (2006). Physiological and genetical aspects of the broad based potato population (Solanum tuberosum L.) in the Netherlands and Northern Finland. *Suomen maataloustieteellisen seuran tiedote* , **21**, 1-7.

Analyzing veterinary surveillance data: Approaches to model the relationship between disease incidence and cattle trade

Birgit Schrödle[1], Leonhard Held[1], Michaela Paul[1]

[1] University of Zurich, Institute for Social and Preventive Medicine, Hirschengraben 84, 8001 Zurich, Switzerland, Email:birgit.schroedle@ifspm.uzh.ch

Abstract: Two approaches to the analysis of registry data for bovine diseases with regard to the relationship between disease incidence and cattle trade are proposed. Firstly, a parameter-driven spatio-temporal disease mapping model formulated in a hierarchical Bayesian framework is used. Various cattle movement parameters, e.g. the number and proportion of in-movements from infected regions, can be included as potential covariates. Within this context problems of such an endogenous covariate are discussed. Since a purely parameter-driven approach is often not adequate to depict local epidemics, a so-called observation-driven infectious disease model is proposed as a second possibility. It includes an autoregressive part for counts in the region of interest in the past. Additionally, the sum of previous cases in other regions weighted by cattle movements is added to assess the spread of the disease by trading. Both models are applied to cases of Coxiellosis in Switzerland, 2005 to 2009.

Keywords: Cattle trade; Spatio-temporal disease mapping; Infectious disease; INLA; Observation-driven

1 Introduction

The spread of a bovine disease can take place over short distances between adjacent or nearby farms borne by wind or insects, which typically results in local clustering of cases. However, disease dispersal also takes place over long distances caused by trade of infectious animals (Gilbert et al., 2005). Hence, the inclusion of cattle trade in an analysis of veterinary surveillance data might give hints towards the association of animal movement and disease presence.

As a first approach disease counts can be analyzed using a disease mapping model that considers spatial, temporal, and spatio-temporal trends (Knorr-Held, 2000; Schrödle and Held, 2009). Additionally, various cattle movement parameters can be included in the model using ecological regression (Clayton et al., 1993). Bayesian inference is conducted using integrated nested Laplace approximations (INLA), which was recently proposed in Rue et al. (2009).

As an alternative to this purely parameter-driven approach the spread of a bovine disease by cattle trade can be modelled within a likelihood-based infectious disease framework (Paul et al., 2008). Here, an autoregressive term for past counts in the region of interest and the sum of past cases in other regions weighted by cattle movements are part of the model formulation. The advantage of this so-called observation-driven model is that it is able to describe local epidemics.
Coxiellosis in cattle is an infectious, bacterial disease among ruminant animals, which can be spread by airborne infection. It can be the reason for an abortion, even in a late phase of the pregnancy. Cases of Coxiellosis were reported to the Swiss Federal Veterinary Office (BVET) between 2005 and 2009 and are available aggregated for 184 Swiss regions and the Principality of Liechtenstein on a yearly basis. The number of herds m_i in each region i is known.
Since 2008 it has been mandatory for Swiss stock-keepers to notify the BVET of all cattle movements. As the spatial pattern of movements is similar for 2008 and 2009 the considered movement parameters are assumed to be consistent from year to year.

2 Spatio-temporal disease mapping

For modelling spatio-temporal disease counts $y_{it} \sim \text{Po}(\exp(\eta_{it}))$ a nonparametric hierarchical Bayesian setting as proposed in Knorr-Held (2000) is used. The respective linear predictor can be written as

$$\eta_{it} = \log(m_i) + \xi + \nu_i + \psi_i + \gamma_t + \phi_t + \delta_{it}, \tag{1}$$

where ξ is an intercept, ν_i and ψ_i are spatially unstructured and structured effects and γ_t and ϕ_t are temporal main effects, specified as an i.i.d. term and a random walk of first order, respectively. The term δ_{it} accounts for spatio-temporal interaction and can be specified assuming four different types of interaction between time and space (Knorr-Held, 2000). To account for covariates x_{it} as introduced in the following paragraph (1) can be extended to

$$\eta_{it} = \log(m_i) + \xi + \nu_i + \psi_i + \gamma_t + \phi_t + \delta_{it} + \beta \cdot x_{it}. \tag{2}$$

The total number of in-movements (model acronym: TOT) and the absolute number (A) and proportion (P) of in-movements from regions with elevated risk are considered as potentially associated with disease presence. Since we assume that the movement pattern does not change from year to year the total number of in-movements is constant over time. The movements from infected regions are time-varying and defined using a two-stage process: A separate spatial disease mapping model is fitted for each year, including only ν_i and ψ_i from (1) (Besag et al., 1991). Regions are indicated

as infected when exceeding two different thresholds, namely an estimated relative risk larger than 2 and 3, respectively. At the second stage the models are fitted using a time lag of one and two years, respectively, to detect the incubation period of the disease (TL1 and TL2). As the time-varying covariate is derived using previous observations it is a so-called endogenous or feedback variable. This issue will be discussed briefly in Section 4.
All models are fit using integrated nested Laplace approximations (INLA). This approach for approximate Bayesian inference was recently proposed by Rue et al. (2009) as an alternative to Markov chain Monte Carlo mechanisms. The advantage of INLA is that it runs in remarkably fast computational time and returns accurate parameter estimates for a wide range of models. Additionally, INLA computes the deviance information criterion (DIC) as tool for Bayesian model choice. All analyses in this paper were conducted using the `R INLA` package build on the 1st of February 2010, version 1.668.

3 Infectious disease model

A purely parameter-driven model as proposed in Section 2 might not be able to describe localized epidemics which can often be found in veterinary disease surveillance data (Held et al., 2005). Hence, a so-called observation-driven model is built including the number of cases $y_{i,t-1}$ in the past. In its simplest formulation the observations y_{it} are Poisson distributed with mean

$$\mu_{it} = \lambda \cdot y_{i,t-1} + m_i \cdot \exp(\alpha) \tag{3}$$

and $\lambda > 0$ (Held et al., 2005). The parameter α accounts for all residual variation. Cases at times $t - k$, $k > 1$, could be considered as well.
As an addition, the sum of counts in all other regions j weighted by a factor w_{ji} can be added to model the spatial spread of the disease over time. In Paul et al. (2008) the respective mean is specified as

$$\mu_{it} = \lambda \cdot y_{i,t-1} + \rho \cdot \sum_{j \neq i} w_{ji} \cdot y_{j,t-1} + m_i \cdot \exp(\alpha) \tag{4}$$

with λ, $\rho > 0$. To assess the association between cattle movement and disease presence the square root of the absolute number of cattle movements (CM) between regions j and i are used as weights w_{ji} in this application. Other weights w_{ji} can also be considered (Paul et al., 2008). Here, models with $w_{ji} = 1$ for all j and for all $j \sim i$, respectively, are fit as alternatives. The term $j \sim i$ denotes all regions j which are neighbours of region i.
The parameter α in (4) can be split into an intercept and a linear time trend

$$\mu_{it} = \lambda \cdot y_{i,t-1} + \rho \cdot \sum_{j \neq i} w_{ji} \cdot y_{j,t-1} + m_i \cdot \exp(\alpha + \zeta \cdot t). \tag{5}$$

TABLE 1. Spatio-temporal disease mapping (see Section 2): The DIC and the posterior mean of the respective cattle trade parameter along with its 95%-credible interval are shown for each model. For the model without covariate (1) the DIC is shown only.

		RR> 2		RR> 3	
		DIC	$\hat{\beta}$	DIC	$\hat{\beta}$
TL1	A	939.3	$0.017\,[0.008; 0.027]$	942.7	$0.012\,[0.004; 0.020]$
	P	947.0	$1.29\,[0.39; 2.18]$	950.0	$0.96\,[-0.02; 1.94]$
TL2	A	945.2	$0.009\,[-0.000; 0.018]$	951.0	$0.007\,[-0.003; 0.016]$
	P	948.0	$0.83\,[-0.038; 1.69]$	951.1	$0.85\,[-0.20; 1.90]$
TOT		955.8	$0.005\,[-0.006; 0.017]$		
(1)		954.6			

Extensions for a region-specific random effect are also possible (Paul et al., 2009), but the computation of the results might suffer from numerical problems if the number of regions is large. Hence, they are not considered here.

Maximum likelihood inference is performed using iterative algorithms as described in Held et al. (2005) and Paul et al. (2008). The AIC is calculated for model choice.

4 Results

Regarding spatio-temporal disease mapping, model (1) was run without covariate for all four possible types of space-time interaction. The model including an interaction term of Type II had the lowest DIC and was chosen as basis model for the ecological regression including cattle trade quantities. All results are summarized in Table 1. The DIC is lowest for the model including the absolute number of in-movements from infected regions with relative risk larger than 2 and a time lag of one year. This is plausible considering the nature of the disease. Models including the absolute number of in-movements are generally preferred compared to models involving the proportion of cattle trade from infected regions. A positive association is obtained for all covariates. For three of the models with a time lag of one year the 95%-credible interval includes only positive values.

As noted in Section 2 the number of in-movements from infected regions is an endogenous covariate. If the pattern of the disease exhibits local clusters, the infected areas chosen by the two-stage process typically are a few groups of neighboring regions. As cattle trade is much larger between neighboring regions, the respective parameters might just explain parts of the local spatial clustering of cases in the data. One hint pointing in this direction is the fact that the estimated variance of the spatially structured effect ψ_i in

TABLE 2. Infectious disease model (see Section 3): The AIC and the estimated coefficients along with their standard errors are shown for each model.

	w_{ji}	AIC	$\hat{\lambda}$	$\hat{\rho}$	$\hat{\zeta}$
(3)	–	1092.8	0.44 (0.05)		
(4)	1	1094.8	0.44 (0.05)	0.0000 (0.0000)	
(4)	1, if $j \sim i$	1062.7	0.43 (0.05)	0.0491 (0.0103)	
(4)	$\sqrt{\text{CM}}$	1082.0	0.44 (0.05)	0.0005 (0.0002)	
(5)	$\sqrt{\text{CM}}$	1083.5	0.44 (0.05)	0.0005 (0.0002)	0.08 (0.11)

(1) drops after inclusion of cattle movement in the model. Unfortunately, it cannot be quantified to what extent such confounding is present.
Results for the infectious disease models are shown in Table 2. With regard to AIC model (4) using $\sqrt{\text{CM}}$ as weights performs better than model (3). Hence, the autoregressive inclusion of counts from other regions weighted by cattle trade provides a better fit. The respective parameter estimate $\hat{\rho}$ is positive (0.0005) with a small standard error (0.0002). In contrast, the alternative model with $w_{ji} = 1$ for all j is not better than (3). If the counts in neighbouring regions $j \sim i$ are considered as additional explanatory variables the AIC is even smaller than for the model including cattle trade and the estimated coefficient is significantly positive (0.0491). Hence, a high local clustering of cases is present. For (5) a positive linear time trend is estimated (0.08), but it is not significantly different from zero.

5 Discussion

Two very different approaches were applied to data on Coxiellosis in cattle to assess the spread of the disease by cattle trade. In both cases model choice criteria and estimated coefficients indicate a positive association between animal movement and disease presence. Nevertheless, both approaches are not without problems. The disease mapping approach makes use of an endogenous covariate which might result in a confounding problem. With regard to the infectious disease model first steps in the direction of a mixture of the parameter- and observation-driven approach are taken when fitting (5). Nevertheless, it would be desirable to include spatial effects as well, especially a spatially structured effect to account for local clustering of the disease. Hence, it must be explored if both approaches could be combined by substituting α in (3) by (1) (without $(\log m_i)$). In this new setting parameter estimation might be possible in a Bayesian framework as algorithms used for maximum likelihood inference (Paul et al., 2008) will possibly suffer from numerical problems. Furthermore, the two approaches are not comparable at the moment as different model choice criteria are derived and different components are included.

Acknowledgments: Financial support by the Swiss Federal Veterinary Office (BVET) is gratefully acknowledged.

References

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

Clayton, D., Bernardinelli, L. and Montomoli, C. (1993). Spatial correlation in ecological analysis. *International Journal of Epidemiology*, **22**, 1193-1202.

Gilbert, M., Mitchell, A., Bourn, D., Mawdsley, J., Clifton-Hadley, R. and Wint, W. (2005). Cattle movement and bovine tuberculosis in Great Britain. *Nature*, **435**, 491-496.

Held, L., Höhle, M. and Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling*, **5**, 187-199.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, **19**, 2555-2568.

Paul, M., Held, L. and Toschke, A. (2008). Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine*, **27**, 6250-6267.

Paul, M. and Held, L. (2009). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Technical Report, University of Zurich.*

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *Journal of the Royal Statistical Society, Series B*, **71**, 319-392.

Schrödle, B. and Held, L. (2009). Spatio-temporal disease mapping using INLA. *Technical Report, University of Zurich.*

Functional Clustering based on Multiresolution Warping

Leen Slaets [1], Gerda Claeskens [1]

[1] ORSTAT and Leuven Statistics Research Center, Katholieke Universiteit Leuven, Naamsestraat 69, 3000 Leuven, Belgium. E-mail: leen.slaets@econ.kuleuven.be and gerda.claeskens@econ.kuleuven.be

Abstract: Typical in a dataset with functional observations is the presence of phase variability in addition to amplitude variability. In this paper we develop a multiresolution warping method which eliminates the phase variation and furthermore efficiently uses the results of the warping step (often overlooked) in a cluster analysis of functional data.

Keywords: Functional Data; Clustering; Time Warping; Curve Registration.

1 Functional Data

In a functional data analysis the observations come from underlying (smooth) functions. Often, the dependent variable is referred to as time t, since many functional data sets arise from observing a process in a certain time interval. Ramsay and Silverman (2006) provide an overview of the statistical methodology to analyze this type of data. For many reasons, the rate at which data points are observed over time varies and does not necessarily reflect the underlying biological, physical, or other process governing the data. There can also be variation in the amplitude of the important curve features (such as peaks and valleys). These two sources of variability, as displayed in Figure 1 are, respectively, phase and amplitude variability.

2 Multiresolution Time warping

2.1 Warping Components

Time warping aims at reducing phase variability present in a sample of functional curve observations. This is usually done by applying continuous strictly increasing functions, the so called warping functions, on the function argument. Popular multiresolution approaches such as splines have been successfully applied to model warping functions (e.g. Gervini and Gasser, 2005) by applying some restrictions to guarantee monotonicity. However, splines are no warping functions and hence the decomposition structure does not have a meaningful interpretation in the context of warping. In Claeskens, Silverman and Slaets (2010) we propose a natural representation of warping functions by composing a novel type of elementary

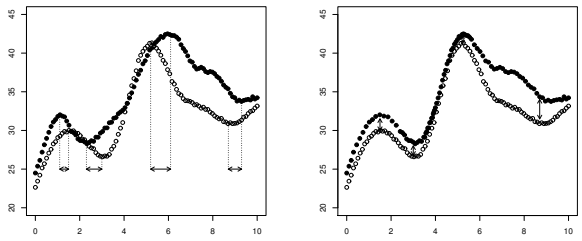


FIGURE 1. Two curve observations with both phase and amplitude variability (left) and the same curves after the elimination of the phase variation (right).

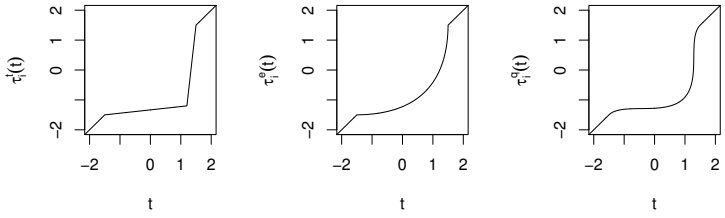


FIGURE 2. Warping components for a triangular, Epanechnikov and quartic kernel.

functions named warping components, τ_i. The latter are rescaled kernel functions, rotated alongside the first diagonal (Figure 2), resulting in warping functions localized in location and scale. Each component is characterized by four parameters: a lower bound w_l, upper bound w_u, center a and an intensity λ. The warping function τ_i can be easily inverted as follows: for $\tau_i(t) = \tau_{i,S}(w_{l,S}, w_{u,S}, a_S, \lambda_S; t) \circ \ldots \circ \tau_{i,1}(w_{l,1}, w_{u,1}, a_1, \lambda_1; t)$ we have $\tau_i^{-1}(t) = \tau_{i,1}(w_{l,1}, w_{u,1}, a_1, -\lambda_1; t) \circ \ldots \circ \tau_{i,S}(w_{l,S}, w_{u,S}, a_S, -\lambda_S; t)$.

2.2 Model Formulation

The data are assumed to be observations $y_i(t_j) = y_{i,j}$ of random curves that are noisy versions of N unobserved continuous curves $(t, F_i(t))$ on a fixed set of time points $t_{i,j}$, $j = 1, \ldots, T$ (for simplicity of notation), for $i = 1, \ldots, N$. The curves $(t, F_i(t))$ originate from an unknown curve μ after warping the time domain and adding an amplitude component:

$$y_{i,j} = F_i(t_{i,j}) + e_{i,j} = \mu(\tau_i(t_{i,j})) + \sum_{k=1}^{K} \beta_{i,k}\phi_k(\tau_i(t_{i,j})) + e_{i,j},, \qquad (1)$$

with $\beta_{i,k}$ and $e_{i,j}$ independent realizations of respectively $\mathcal{N}(0, \sigma_k^2)$ and $\mathcal{N}(0, \sigma^2)$ for $i = 1, \ldots, N$, $j = 1, \ldots, T, k = 1, \ldots, K$, and where $\{\phi_k\}_{k=1}^K$ are asymmetric quartic kernel functions to capture amplitude variation. A sum of weighted pairwise log-likelihoods is used to estimate the parameters in the warping functions τ_i, the variances of the random amplitudes and error terms and the center, lower and upper bound of each kernel. Additionally we allow for an initial shift of the curves in each of the warping functions τ_i. The number of kernels K is specified by the user. For the estimation and the selection of the number of warping components a special Bayesian strategy was introduced with the MCMC computations implemented in C++.

2.3 Simulation Results

One hundred datasets were generated consisting of 5 curves each, in which both phase and amplitude variability are present. The performance of multiresolution warping (with $\{\phi_k\}_{k=1}^K$ in (1) user-defined B-spline basis functions instead of kernels) is compared to that of the popular continuous monotone registration (Ramsay and Li ,1997) and the more advanced nonparametric maximum likelihood registration (Gervini and Gasser, 2005). Several criteria measure the quality of the warp by looking at the alignment of landmarks and the elimination of amplitude variability in between those landmarks. Overall the multiresolution method is most successful in meeting these two requirements. For more details we refer to Slaets, Cleaskens and Silverman (2010).

3 Clustering

In many functional data sets one can observe groups of curves characterized by a similar pattern or similar features and clustering techniques are used to identify them. Classic multivariate methods, such k-means clustering and partitioning around medoids (PAM), can in principle be applied, but they do not take into account the functional nature of the data. Modification of these algorithms specifically designed for functional data have already been proposed in literature, by introducing novel dissimilarity measures (e.g. Chiou and Li, 2007). The additional problem of phase variation in the functional data framework is however often left unaddressed, even though variation in timing of features could be the most important trait of clusters. Therefore it is desirable to incorporate the time warping stage into the clustering analysis. While James (2009) combined the warping functions and the warped curves in a functional principle components analysis, we can rely on the meaningful dimension reduction of the warping function. In the parameterization of the warping functions we only allow for curve specific intensity parameters in each component, which will characterize the

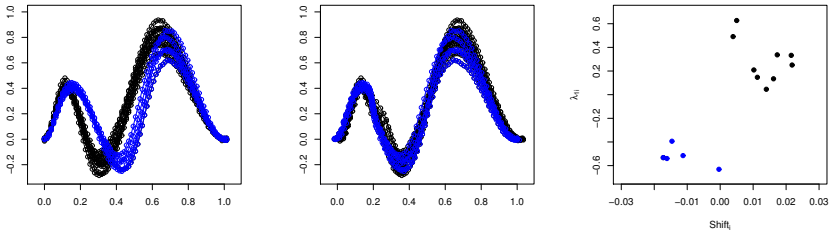


FIGURE 3. Orginal data (left), curves after multiresolution warping with shift and 1 warping components (middle) and scatterplot of the estimated shifts and curve specific intensities $\lambda_{1,i}$ (right).

phase for each curve. The following example illustrates this idea. Figure 3 shows a sample of 14 simulated curves in which the four blue curves belong to a separate cluster with respect to their phase. After warping, the robust PAM algorithm was applied on the estimated shifts and intensities $\lambda_{1,i}$ $(i = 1, \ldots, 8)$ and identified the correct curves as a seperate cluster. We will also further include amplitude variability and compare our method with the most advanced and popular methods.

References

Chiou, J.-M., and Li, P.-L. (2007). Functional Clustering and Identifying Substructures of Longitudinal Data. *Journal of the Royal Statistical Society, Series B.* **69**, 679-699.

Claeskens, G., Silverman, B.W. and Slaets, L. (2010). A Multiresolution Approach to Time Warping Achieved by a Bayesian Prior-Posterior Transfer Fitting Strategy. *Journal of the Royal Statistical Society, Series B.* (under revision).

Gervini, D., and Gasser, T. (2005). Nonparametric Maximum Likelihood Estimation of the Structural Mean of a Sample of Curves.

James G. (2009). Moments Based Functional Synchronization. (unpublished manuscript)

Ramsay, J.O. and Li, X. (1998). Curve Registration. *Journal of the Royal Statistical Society, Series B.* **60**, 351-363.

Ramsay, J.O., and Silverman, B.W. (2006). *Functional Data Analysis.* New York: Springer.

Asymmetry in Breast Reconstruction Patients

Joanna Smith[1], Adrian Bowman[1]

[1] Department of Statistics, University of Glasgow, G12 8QQ.

Keywords: shape analysis; asymmetry; landmarks; breast reconstruction.

1 Background

There is interest in knowing the extent of asymmetry present in the breasts of patients who have undergone a unilateral mastectomy and reconstruction procedure. It is widely accepted that there is a natural breast asymmetry present in the wider population (as studied in Losken et al (2005) and Brown et al (1999)) and interest lies in whether there is a more pronounced asymmetry within this patient group than would normally be seen, and where this asymmetry lies. An analysis was carried out on 44 women between the ages of 37 and 67, all having undergone this procedure. Three-dimensional images were captured for each patient using a 3D stereo-photogrammetry system, and each case was then marked with ten anatomically significant landmarks. The data were collected as part of a collaborative study with the Glasgow Dental Hospital.

2 Methods

2.1 Landmark Asymmetry

Asymmetry can be quantified as the degree to which there is a mismatch between a landmark configuration (the set of all landmarks on an individual image) and its relabelled and matched reflection, as discussed in Bock and Bowman (2006). Several of the landmarks have natural pairings due to their corresponding positions on the left and right breasts, whilst others lie on the midline of the chest. When reflected, the paired points must be matched to each other, and the midline points must be matched to themselves. After reflecting, rotating and scaling to minimise sums of squares distances between corresponding landmarks we should have removed any location, orientation and size effects and be left purely with the genuine shape differences. This can be quantified into an asymmetry score for each

patient. As the scores arise from sums of squares, a square-root transformation can be applied to reduce skewness. To check the validity of the results, subjective evaluations of symmetry were also obtained.
While the previously calculated asymmetry scores give an indication of the overall asymmetry present in a case, it is possible to examine what factors are contributing to this asymmetry as well. As discussed in Bock and Bowman (2006), if the rest of the surface was held fixed and the reconstructed breast was translated, rotated or scaled the overall asymmetry score would change, and we can use this to assess how much of the asymmetry that is present is due to the location, orientation and size of the reconstructed breast. It follows that any asymmetry remaining after these transformations is due to a difference in the actual shape of the breasts, or an 'intrinsic asymmetry'.

2.2 Surface Asymmetry

It is also desirable to examine asymmetry over the whole surface of the breasts rather than just the landmarks. The landmarks give us some idea of shape but are only a very small set of points on a surface composed of many thousands of points, therefore we are only using a fraction of the information available to us. In order to do this it is necessary to decide which points to include as 'breast', and which should be treated as chest wall and disregarded. This can be done by calculating the curvature and using it to define the breast boundary. After the extraction of the breasts from the image, we can begin to examine the asymmetry of the surfaces. In order to do this, we create a point distribution model. This creates a set of corresponding points across all breasts, to ensure that they all have the same number of points which are in corresponding positions. Then, after reflection, the asymmetry can be quantified by calculating the distances between these corresponding points on the reconstructed and unreconstructed breast. The shape differences between the two breasts can also be examined by a principal components analysis.

3 Results

The average asymmetry score found from the landmarks was found to be 0.052, with scores ranging from 0.019 to 0.136. Figure 1 shows the images and configurations for the cases with the lowest and highest asymmetry scores respectively.It was found that the relationship between the asymmetry scores and subjective scores was highly significant ($p < 0.0001$), with a correlation coefficient of -0.62. As would be expected it is a negative relationship, as an increase in asymmetry leads to a decrease in subjective score.

Decomposing the asymmetry score into its various components, the location of the reconstructed breast on the chest wall contributed the most significantly to asymmetry (34.6%), with the intrinsic asymmetry (genuine shape differences between the breast being the second most influential component (35.2%). Orientation and scaling contributed 18.8% and 11.3% respectively. The overall pattern and individual results are shown in Figure 2.

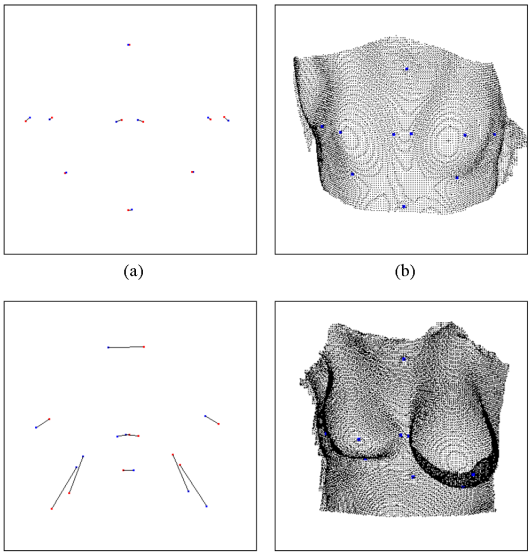


FIGURE 1. (a) Result of reflecting and matching in the case with the lowest asymmetry score, with both the original landmarks (blue) and the reflected (red) (b) The patient with the lowest asymmetry score ($A = 0.019$) (c) Result of reflecting and matching in the case with the highest asymmetry score (d) The patient with the highest asymmetry score ($A = 0.136$).

Similar results were found from the scores calculated over the whole breast surface. The breakdown of the scores was very similar to the landmark results, with the location of the reconstructed breast accounting for 36.6%, intrinsic asymmetry for 31.8%, orientation for 19.8% and finally scaling for 11.8%. These scores were also found to be have a highly significant relationship with the subjective scores ($p < 0.001$), but the correlation was slightly weaker than with the landmark results (-0.49). This method relies on finding the correct boundary, so this could be due to errors in this area.

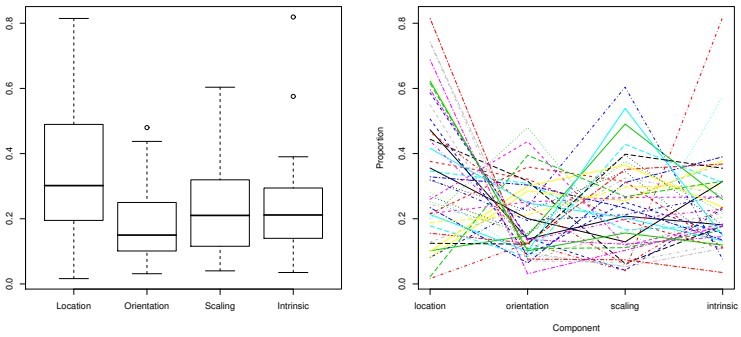


FIGURE 2. (a) Boxplot showing the proportions of asymmetry accounted for by individual components of asymmetry (b) Proportion of asymmetry accounted for by individual components of asymmetry in each case

Acknowledgments: The breast images and landmark data were kindly provided by Dr. Helga Henseler, Canniesburn Plastic Surgery Unit, Glasgow Royal Infirmary.

References

Bock, M.T. and Bowman, A.W. (2006). On the measurement and analysis of asymmetry with applications to facial modelling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **55**, 77–91.

Losken, A., Fishman, I., Denson, D.D., Moyer, H.R. and Carlson, G (2005) An Objectice Evaluation of Breast Symmetry and Shape Differences Using 3- Dimensional Images. *Annals of Plastic Surgery*, **55**, 571–575.

Brown, T.P., Ringrose, C., Hyland, R.E., Cole, A.A. and Brotherston, T.M. (1999). A method of assessing female breast morphometry and its clinical application. *British Journal of Plastic Surgery*, **52**, 355–359.

Spatial small area estimation: a comparison of Non-parametric EBLUP and M-quantile GWR models

Georgy Yu. Sofronov[1]

[1] Department of Statistics, Macquarie University NSW 2109 Australia. E-mail: georgy.sofronov@mq.edu.au

Abstract: In this paper we compare the performance of two spatial small area estimation (SAE) methods — the Non-parametric Empirical Best Linear Unbiased Predictor (NPEBLUP) and the M-quantile Geographically Weighted Regression (MQGWR) model. The Root Mean Squared Error (RMSE) is computed as a measure of estimation performance of the predictors. The properties of the estimators are evaluated by applying them to the results of farm surveys that have been conducted by the Australian Bureau of Agricultural and Resource Economics.

Keywords: Small Area Estimation; Spatial Models; Non-parametric Empirical Best Linear Unbiased Predictor; M-quantile Geographically Weighted Regression model.

1 Spatial models for small area estimation

Estimation of population characteristics for sub-national domains (or smaller regions) is an important objective for statistical surveys. In particular, geographically defined domains, e.g. regions, states, counties, wards and metropolitan areas can be of interest. In various applications, observations that are spatially close may be more related than observations that are further apart. Methods that incorporate the spatial information in a regression model enable researches to borrow strength over space and hence potentially improve the precision of small area estimates.

In what follows we assume that a vector of p auxiliary variables x is known for each population unit j in small area i and that information for the variable of interest y is available from a sample which we denote by s that includes units from all d small areas of interest. We denote the population (sample) size in area i by $N_i(n_i)$ and use $s_i(r_i)$ to denote the sampled (non-sampled) population units in this area. The overall population (sample) size is $N(n)$. The target is to use these data to estimate various area specific quantities, including (but not only) the small area mean m_i of y.

1.1 Non-parametric Empirical Best Linear Unbiased Predictor

A flexible and popular method in small area estimation (SAE) is the use of linear mixed models with area specific random effects, with estimation and inferences typically carried out using empirical best linear unbiased prediction (EBLUP — see Rao (2003)). In the general case a linear mixed model for the value of y in area i has the following form

$$y_{ij} = x_{ij}^T\beta + \gamma_i + \varepsilon_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, d, \tag{1}$$

where ε_{ij} is an individual random effect with mean zero and variance σ_ε^2, γ_i is the random area effect associated with small area i, assumed to have mean zero and variance σ_γ^2. These two terms are assumed to be mutually independent. The prediction of non-observed y_{ij} values is given by $x_{ij}^T\hat{\beta}+\hat{\gamma}_i$, with $\hat{\beta}$ parameter estimate and $\hat{\gamma}_i$ predictor obtained through either maximum likelihood or restricted maximum likelihood procedures. In matrix notation, (1) can be expressed as follows

$$\begin{aligned}
&Y = X\beta + Z\gamma + \varepsilon,\\
&Y = (Y_1^T, \ldots, Y_d^T)^T, \quad Y_i = (y_{i1}, \ldots, y_{in_i})^T,\\
&X = (X_1^T, \ldots, X_d^T)^T, \quad X_i = (x_{i1}, \ldots, x_{in_i})^T,\\
&Z = \operatorname{diag}(Z_i = 1_{n_i}; 1 \le i \le d), \quad \gamma = (\gamma_1, \ldots, \gamma_d)^T,\\
&\varepsilon = (\varepsilon_1^T, \ldots, \varepsilon_d^T)^T, \quad \varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})^T,
\end{aligned}$$

where 1_{n_i} is the unit vector of length n_i.

One popular method in SAE is the Non-parametric Empirical Best Linear Unbiased Predictor (NPEBLUP) (see Opsomer *et al.* (2008)) based on linear mixed models with area specific random effects and a smooth, non-parametrically specified trend. The model has the form

$$Y = X\beta + D\delta + Z\gamma + \varepsilon,$$

where $X\beta$ is a fixed effect, $D\delta$ is a spline portion of a random effect, $\delta \sim \mathsf{N}(0, \Sigma_\delta)$, $\Sigma_\delta = \sigma_\delta^2 I_K$, I_K is the identity matrix of order K, K is the number of knots, $Z\gamma$ is a small area random effect, $\gamma \sim \mathsf{N}(0, \Sigma_\gamma)$, $\Sigma_\gamma = \sigma_\gamma^2 I_d$, ε is an individual random effect, $\varepsilon \sim \mathsf{N}(0, \Sigma_\varepsilon)$, $\Sigma_\varepsilon = \sigma_\varepsilon^2 I_n$. Each of the random components is assumed to be independent of the others. The NPEBLUP of m_i is of the form

$$\begin{aligned}
&\hat{m}_i = \bar{x}_i\hat{\beta} + \bar{d}_i\hat{\delta} + b_i^T\hat{\gamma},\\
&\hat{\delta} = \hat{\sigma}_\delta^2 D^T\widehat{V}^{-1}(Y - X^T\hat{\beta}), \quad \hat{\gamma} = \hat{\sigma}_\gamma^2 Z^T\widehat{V}^{-1}(Y - X^T\hat{\beta}),\\
&\hat{\beta} = (X^T\widehat{V}^{-1}X)^{-1}X^T\widehat{V}^{-1}Y, \quad \widehat{V} = \hat{\sigma}_\delta^2 DD^T + \hat{\sigma}_\gamma^2 ZZ^T + \hat{\sigma}_\varepsilon^2 I_n,
\end{aligned}$$

where $\hat{\sigma}_\delta^2$, $\hat{\sigma}_\gamma^2$, $\hat{\sigma}_\varepsilon^2$ are estimates of σ_δ^2, σ_γ^2, σ_ε^2, respectively, b_i is the d-vector $(0, 0, 0, \ldots, 1, \ldots, 0, 0)^T$ with the 1 in the ith position, $\bar{x}_i$ and $\bar{d}_i$ are the true means of the powers of x_i (up to the degree of the spline) and of the spline basis functions over the small area. Both $\bar{x}_i$ and $\bar{d}_i$ are assumed to be known.

1.2 M-quantile Geographically Weighted Regression Models

An alternative approach to SAE is based on the use of M-quantile models (Breckling and Chambers (1988)). The M-quantile of order q of a random variable Y with distribution function $F(y)$ is the value m_q that satisfies

$$\int \psi_q\left(\frac{y-m_q}{\sigma_q}\right) dF(y) = 0,$$
$$\psi_q(u) = \{(1-q)I(u<0) + qI(u \geq 0)\}\,\psi(u)$$

and ψ is an appropriately chosen influence function. Here $\sigma_q = E|Y - m_q|$ is a measure of the scale of the residuals from the M-quantile m_q. Note that when $\psi(u) = \text{sgn}(u)$ we obtain the standard quantile of order q.
Breckling and Chambers (1988) define a linear M-quantile regression model is one where the qth M-quantile $Q_q(X;\psi)$ of the conditional distribution of y given x satisfies

$$Q_q(x_{ij};\psi) = x_{ij}^T \beta_\psi(q). \tag{2}$$

For specified q and continuous ψ, an estimate $\hat{\beta}_\psi(q)$ of $\beta_\psi(q)$ can be obtained via an iterative weighted least squares algorithm.
It can be defined a spatial extension to linear M-quantile regression based on Geographically Weighted Regression (GWR) (see Fotheringham *et al.* (2002)). Given n observations at a set of L locations $\{u_l; l = 1,\ldots,L; L \leq n\}$, a model for the M-quantile of order q of the conditional distribution of y given x at u by allowing (2) to depend on u. That is, we write

$$Q_q(X;\psi,u) = X^T \beta_\psi(u;q), \tag{3}$$

where $\beta_\psi(u;q)$ varies with u as well as with q. That is, (3) allows the entire conditional distribution (not just the mean) of y given x to vary from location to location.
Following Chambers and Tzavidis (2006), an M-quantile GWR predictor of the mean m_i in small area i is then

$$\hat{m}_i = N_i^{-1}\left(\sum_{j\in s_i \cup r_i} \hat{Q}_{\hat{\theta}_i}(x_j;\psi,u_j) + \frac{N_i}{n_i}\sum_{j\in s_i}\left(y_j - \hat{Q}_{\hat{\theta}_i}(x_j;\psi,u_j)\right)\right)$$

where $\hat{Q}_{\hat{\theta}_i}(x_i;\psi,u_j)$ is defined via the MQGWR model (3), $\hat{\theta}_i$ is the average value of the sample MQGWR coefficients in area i. The coefficient for unit j with values y_j and x_j at location u_j is the unique value q_j such that $\hat{Q}_{q_j}(x_j;\psi,u_j) = y_j$.

1.3 Mean squared error estimation

The Mean Squared Error (MSE) estimator of both predictors can be found by using the pseudo-linearization approach to MSE estimation that described in Chambers *et al.* (2008). The basic idea of this approach is to

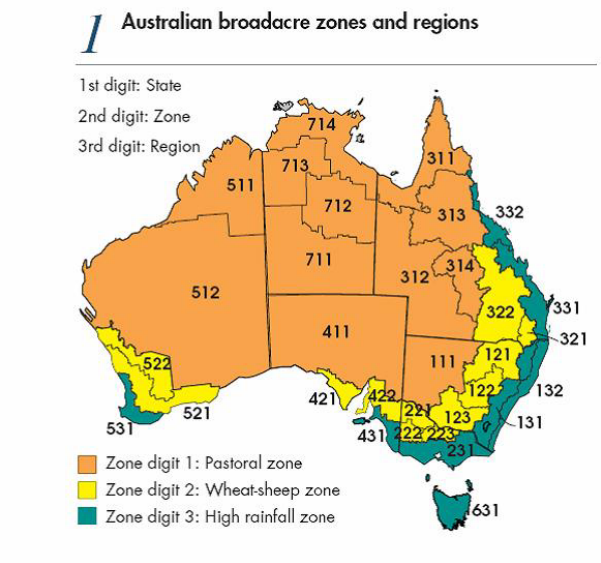


FIGURE 1. Australian broadacre zones and regions.

express a predictor in pseudo-linear form, i.e. as a weighted sum of the sample values of y:

$$\hat{m}_i = \sum_{j \in s} w_{ij} y_j,$$

and applying robust MSE estimation methods that treat these weights (which typically depend on estimated variance components) as fixed.

2 Data Analysis

We compare the performance of two SAE methods — the NPEBLUP and the MQGWR model. We also use two direct estimates — Direct(1) (sample weighted average) and Direct(2) (unweighted average). The Root Mean Squared Error (RMSE) is computed as a measure of estimation performance of the predictors. The properties of the estimators are evaluated by applying them to the results of farm surveys that have been conducted by the Australian Bureau of Agricultural and Resource Economics. For testing of the small area models we use unit record farm level survey data from the wheat-sheep zone (see Figure 1) for the survey in year 1990. Synthetic coordinates (longitude, latitude) for each farm were obtained by using pairwise distances between farms. The variable of interest is average Total Cash Receipts (TCR) within 12 small areas.

The fixed effect part of the NPEBLUP(1) model includes the following variables: land area, four dummy variables made up of five industries (spe-

TABLE 1. Predictor for the ith small area mean.

i	NEBLUP(1)	NEBLUP(2)	MQGWR	Direct(1)	Direct(2)
1	243728	292115	176984	230367	632086
2	255644	224224	176779	168975	160812
3	227058	255666	206885	184569	169001
4	235366	240796	192632	188289	111925
5	224629	134041	153267	154100	99914
6	172400	172468	120850	108881	72236
7	107979	104153	181257	142112	99589
8	-296079	-88431	219510	225562	187099
9	241560	194796	218406	209798	163649
10	253334	222084	166813	157579	114856
11	339745	174038	259466	277049	322312
12	353163	432578	285678	311246	225047

cialist croppers, mixed livestock croppers, sheep specialists, beef specialists, mixed sheep beef farms), number of closing stock-beef, number of closing stock-sheep, wheat quantity harvested. The non-parametric part of the NPEBLUP(1) model includes matrix D for the spatial locations ($K = 60$). The fixed effect part of the model NPEBLUP(2) includes four dummy variables made up of five industries. The non-parametric part of the model NPEBLUP(2) includes matrix D as for the spatial locations and the following variables: land area, number of closing stock-beef, number of closing stock-sheep, wheat quantity harvested. In the NPEBLUP(2) model we use 14 knots for land area, number of closing stock-beef, number of closing stock-sheep, and wheat quantity harvested. Thus, the parameter $K = 60 + 4 \cdot 14 = 116$ for the NPEBLUP(2).

For both models we use the space filling algorithm that is implemented in the SemiPar package for R.

The MQGWR model includes all the variables used in the fixed part of the NPEBLUP(1) model plus the synthetic locations of each farm in a population data set.

Table 1 shows that both the NPEBLUP(1) and the NPEBLUP(2) can be very unstable in the case of presence of outliers (the predictors may give small or even negative values of average TCR).

The results set out in Table 1 and Table 2 show that the MQGWR model appears to be superior to the other predictors in terms of estimated RMSE and robustness. However the MQGWR is time-consuming since it is required to evaluate parameters in each location in the population. The NPEBLUP(2) has larger estimated RMSE than the NPEBLUP(1) due to larger number of knots. The direct predictor can perform better than other predictors if they are not well specified.

TABLE 2. Estimated Root Mean Squared Error.

i	NPEBLUP(1)	NPEBLUP(2)	MQGWR	Direct(1)	Direct(2)
1	120617	270156	27508	196403	899582
2	130218	143177	28727	60992	55286
3	53190	80772	13616	50930	42336
4	65707	83653	20712	52372	55608
5	90117	77297	15533	49432	49556
6	52745	78341	8932	39217	37779
7	183219	104174	21966	48653	47556
8	252980	145528	16638	52259	51818
9	109775	108984	16520	82020	70678
10	90840	85507	14022	38953	39635
11	156280	106800	10234	88170	192491
12	371809	287330	12292	80487	80778
Mean	139791	130977	17225	69991	135259

References

Breckling, J., and Chambers, R. (1988). M-quantiles. *Biometrika*, **75**, 761-771.

Chambers, R., Chandra, H., and Tzavidis, N. (2008). On Bias-Robust Mean Squared Error Estimation for Linear Predictors for Domains. Working Paper 09-08. Centre for Statistical and Survey Methodology, The University of Wollongong.

Chambers, R., and Tzavidis, N. (2006). M-quantiles models for small area estimation. *Biometrika*, **93**, 255-268.

Fotheringham, A.S., Brunsdon, C. and Charlton, M. (2002). *Geographically Weighted Regression*, West Sussex: John Wiley & Sons.

Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G., and Breidt, F.J. (2008). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society, Series B*, **70**, 265-286.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: Wiley.

Salvati, N., Tzavidis, N., Pratesi, M., and Chambers, R. (2007). Small area estimation via M-quantile geographically weighted regression. CCSR Working Paper 2007-09. Centre for Census and Survey Research, University of Manchester.

Confidence Intervals for Effect Sizes in a Meta-analysis based on Paired Comparisons and Independent Group Data

R.G. Staudte[1]

[1] Department of Mathematics and Statistics, La Trobe University, Melbourne, Vic. Australia 3086. email r.staudte@latrobe.edu.au

Abstract: Morris and DeShon (2002) present methods for finding point estimates of effect sizes in a meta-analysis of studies based on a combination of repeated measures and independent group designs. We build on this approach and use variance stabilizing techniques to find confidence intervals for these effect sizes, using both fixed and random effects models.

Keywords: effect size; fixed effects model; random effects model; repeated measures; variance stabilization.

1 Introduction

Morris and DeShon (2002) delineate the problems facing a researcher who wants to combine results arising from different designs. They also suggest methods for finding estimated weights required by a traditional meta-analysis, but stop short of finding confidence intervals for an overall representative effect size. This can be done by transforming each effect size estimate $\hat{\delta}_k$ to $\kappa_k = \mathcal{K}(\hat{\delta}_k)$ so that the transformed effect κ_k is asymptotically normal with mean $\kappa = \mathcal{K}(\delta)$ and 'known' variance $1/n_k$, where n_k is the sample size in study k. Such a *variance stabilizing transformation*, or VST, often improves the normality of the estimated effects. Thus one can confidently apply the traditional meta-analysis methods to the $\hat{\kappa}_k$'s, obtaining a weighted (with weights $w_k = n_k$, $W = N = \sum_k n_k$) estimator $\hat{\kappa}_w = \sum_k n_k \hat{\kappa}_k / N$ of κ which is asymptotically normal with mean κ and variance N^{-1}. This leads to confidence intervals for κ of the form $\hat{\kappa}_w \pm cN^{-1/2}$. It is easily back-transformed to an interval for δ by applying the inverse transformation $\mathcal{K}^{-1}$ to the interval for κ.

An important example of a VST which we will apply to estimated effect sizes $\hat{\delta}$ arising in this paper is defined by

$$\mathcal{K}(\delta) = \sqrt{2}\, \sinh^{-1}(\delta/\sqrt{2}\,) \;, \tag{1}$$

where $y = \sinh^{-1}(x) = \ln(x + \sqrt{1+x^2}\,)$ is the inverse of the hyperbolic sine function defined by $x = (e^y - e^{-y})/2$. Given a Student-t statistic S_n

having a non-central t distribution with $n-1$ degrees of freedom and non-centrality parameter $\sqrt{n}\,\delta$, the transformed statistic $T_n = \sqrt{n}\,\mathcal{K}(S_n/\sqrt{n})$ has an approximately normal distribution with parameters $\sqrt{n}\,\mathcal{K}(\delta), 1$. The original VST is due to Azorin (1953), but the simpler expression using (1) is derived in Chapter 20 of Kulinskaya, Morgenthaler and Staudte (2008).
In Section 2 we consider only paired comparisons, and for each of the fixed effects model (FEM) and the random effects model (REM) we find *evidence for* a positive effect size, and also confidence intervals for a representative transformed effect. These intervals are then back-transformed onto intervals for the original effect size. The methods are illustrated on the inter-personal skills training studies summarized in Morris and DeShon (2002).
In Section 3 we require only slight modifications to the same methods to carry out a similar analysis for independent group data, and again illustrate them with a detailed analysis on the independent group studies from Morris and DeShon(2002).

2 Meta-analysis for paired comparisons

Given n independent pairs (X_{1i}, X_{2i}), $i = 1, \ldots, n$ with each pair having the bivariate normal distribution with means (μ_1, μ_2), variances (σ_1^2, σ_2^2) and correlation coefficient ρ, all parameters unknown. The differences $D_i = X_{2i} - X_{1i}$ are independent, each normally distributed with unknown mean $\mu_D = \Delta = \mu_2 - \mu_1$ and unknown variance

$$\sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho\ . \tag{2}$$

Let $\bar{D} = \sum_i D_i/n$ and $s_D^2 = \sum_i (D_i - \bar{D})^2/(n-1)$ be the sample mean and variance of the differences. Then inference for the (raw) effect Δ is readily based on the statistic $t_{n-1} = (\bar{D} - 0)/(s_D/\sqrt{n}\,)$. Under our assumptions t_{n-1} has the non-central t distribution with $n-1$ degrees of freedom and non-centrality parameter $\sqrt{n}\,\delta$, where $\delta = \Delta/\sigma_D$ is a scale-free parameter often called the *effect size*.
The above design is often employed in comparing pre- and post-mean scores of subjects receiving a treatment. Then μ_1 is the mean score prior to treatment and μ_2 is the mean score after treatment; it is often called a *repeated measures* design, and then we denote $\delta = \delta_{RM}$ to distinguish it from the effect size δ_{IG} of the next section. To illustrate a meta-analysis of such data, consider the values of $\hat{\delta}_k$ listed in Table 1, to the two decimal places provided in our source.
The corresponding Student-t statistics $t_{n_k} = \sqrt{n}_k\,\hat{\delta}_k$ have respective non-central t distributions with $n_k - 1$ df and non-centrality parameters δ_k. Hence by the discussion following (1), $\hat{\kappa}_k = \mathcal{K}(\hat{\delta}_k)$ is approximately normal with mean $\kappa_k = \mathcal{K}(\delta_k)$ and variance $1/n_k$. These results allow one to compute individual study intervals for $\hat{\kappa}_k$. By applying the inverse transformation $\delta = \mathcal{K}^{-1}(\kappa) = \sqrt{2}\,\sinh(\kappa/\sqrt{2}\,)$ to this interval, one obtains the corresponding interval $[L_{\delta_k}, U_{\delta_k}]$ for δ_k, shown in columns 6,7 of Table 1.

TABLE 1. *These data are selected from the RM rows of Table 3 of Morris and DeShon (2002). Here n_k is the number of subjects in the kth study who were given interpersonal skills training, and δ_k gives the effect size based on pre- and post-training for each study. The Student-t statistics are defined by $t_{n_k} = \sqrt{n_k}\,\hat{\delta}_k$. The remaining columns are defined and explained in the text.*

Study	n_k	$\hat{\delta}_k$	t_{n_k}	T_k	$\hat{\kappa}_k$	L_{δ_k}	U_{δ_k}
1	27	0.84	4.365	4.142	0.797	0.426	1.314
2	27	0.57	2.962	2.887	0.556	0.179	1.002
3	12	1.40	4.850	4.283	1.236	0.696	2.331
4	12	1.23	4.261	3.852	1.112	0.560	2.100
5	36	3.25	19.500	13.318	2.220	2.511	4.163
6	44	0.94	6.235	5.849	0.882	0.603	1.318
7	35	0.93	5.502	5.167	0.873	0.555	1.356
8	127	0.43	4.846	4.774	0.424	0.251	0.615
9	39	0.13	0.812	0.811	0.130	−0.185	0.451

The transformed test statistic T_k gives the *evidence for* the alternative $\delta_k > 0$. Values of T_k near 1.645 are somewhat arbitrarily defined as 'weak' evidence, near 3.3 as 'moderate' and 5 as 'strong'. One can see that Study 9 has negligible evidence for $\delta_9 > 0$ and all other studies have moderate to strong evidence for positive effect sizes, except Study 5, in which $T_5 = 13.3$ is so extraordinarily large as to be an outlier, worthy of further checking. *We omit this study in our analyses hereafter.*

2.1 Meta-analysis for the fixed effects model

The meta-analysis is carried out on the $K = 8$ remaining transformed effects $\hat{\kappa}_k$, and then back-transformed to the space of the effect sizes $\hat{\delta}_k$. The fixed effects model (FEM) assumes all $\kappa_k = \kappa$. Then the inverse-variance weights approach described in Section 1 applied to our data leads to a total sample size $N = 323$, a 95% confidence interval for κ of $[0.488, 0.706]$. When back-transformed the interval for δ is $[0.498, 0.736]$.

2.2 Meta-analysis for the random effects model

The random effects model (REM) assumes that the κ_k's are a random sample from a normal distribution with mean κ and variance γ^2, and that each conditionally on κ_k, the distribution of $\hat{\kappa}_k$ is asymptotically normal with mean κ_k, variance $1/n_k$. Then the unconditional distribution of $\hat{\kappa}_k$ is asymptotically normal with mean κ and variance $\gamma^2 + 1/n_k$.
If all the sample sizes were equal $n_k = n$, then the $\hat{\kappa}_k$'s would be a random sample of size K from a normal population with unknown mean κ

TABLE 2. *The data in columns 1–4 are selected from the IG rows of Table 3 of Morris and DeShon (2002); statistics are defined in the text.*

k	$n_{T,k}$	$n_{C,k}$	$\hat{\delta}_{IG,k}$	N_k	t_{IG,N_k}	$T_{IG,k}$	$\hat{\kappa}_{IG,k}$	$L_{\delta_{IG,k}}$	$U_{\delta_{IG,k}}$
1	30	30	0.61	60	2.363	2.345	0.303	0.099	1.140
2	30	30	0.89	60	3.447	3.392	0.438	0.371	1.438
3	90	93	0.96	183	6.492	6.374	0.471	0.658	1.272
4	9	9	1.34	18	2.843	2.746	0.647	0.371	2.453
5	190	205	0.26	395	2.582	2.578	0.130	0.062	0.459
6	41	35	0.52	76	2.260	2.247	0.258	0.066	0.987

and unknown variance $\gamma^2 + 1/n$, and one could find the usual Student t interval for κ based on the sample mean $\bar{\kappa} = \sum_k \hat{\kappa}_k/K$ and variance $s_\kappa^2 = \sum_k(\hat{\kappa}_k - \bar{\kappa})^2/(K-1)$. Even when the n_k's are not equal, these t intervals have accurate coverage for all $K > 1$, provided $\gamma > s_{1/n_k}$, where s_{1/n_k} is the sample standard deviation of the reciprocals of the sample sizes. This quantity will often be quite small, so it is only for very small γ that the coverage is overly conservative. See Chapter 25 of Kulinskaya, Morgnethaler and Staudte (2008) for supporting simulation studies.

For the data in Table 1 (minus the outlier) $K = 8$, $\bar{\kappa} = 0.751$ and $s_\kappa = 0.365$. The 95% confidence t-interval for κ is equal to $[0.0446, 1.056]$. After back-transformation by $\mathcal{K}^{-1}(\kappa) = \sqrt{2}\sinh(\kappa/\sqrt{2})$, a 95% interval for δ is $[0.454, 1.157]$. This interval is substantially larger than the one found for the FEM, which was $[0.498, 0.736]$. This is expected given the presence of the unknown variance component γ^2.

Note that is is not necessary to estimate γ^2 to find the above intervals. However one may obtain the DerSimonian and Laird (1986) estimate of it based on Q^*, Cochran's Q applied to the transformed effects. It is $\hat{\gamma}_{DL}^2 = \max\{0, Q^* - (K-1)/(N - \sum n_k^2/N)\}$. For our data $\hat{\gamma}_{DL}^2 = 0.082$, and the estimate of γ is 0.287. This is substantially larger than $s_{1/n_k} = 0.028$. Therefore the coverage is close to 95%.

3 Meta-analysis for independent group designs

The *independent groups design* selects $N = n_T + n_C$ subjects at random. From these subjects n_T are randomly selected to receive a treatment and n_C are selected to serve as controls. A typical treated subject score X_{Ti} is assumed to be normally distributed with mean μ_T, variance σ_T^2, and similarly for each X_{Ci}. Let $\bar{X}_T, s_T^2$ be the sample mean and variance of the treated subject scores, and similarly define $\bar{X}_C, s_C^2$ for the control subject scores. Then $\bar{X}_T$ is normally distributed with mean μ_T, variance σ_T^2/n_T and similarly for the control group.

The inter-group effect $\Delta_{IG} = \mu_T - \mu_C$ is estimated by $\hat{\Delta}_{IG} = \bar{X}_T - \bar{X}_C$. Assuming $\sigma_T^2 = \sigma_C^2 = \sigma^2$, the usual pooled estimate of this common variance is denoted s^2_{pooled}. The inter-group *effect size* is defined by $\delta_{IG} = \Delta_{IG}/\sigma$; it is sometimes called Cohen's d, see Cohen (1988). A standard estimate of it is $\hat{\delta}_{IG} = \hat{\Delta}_{IG}/s_{pooled}$. Let $q = n_T/N$. Then, under our assumptions of independent normal samples, $t_{IG,N} = \sqrt{Nq(1-q)}\,\hat{\delta}_{IG}$, where $t_{IG,N}$ has the non-central t distribution with $N-2$ df and non-centrality parameter $\sqrt{Nq(1-q)}\delta_{IG}$. The evidence for $\delta_{IG} > 0$ is $T_{IG} = \sqrt{N}\,\mathcal{K}(t_{IG,N}/\sqrt{N}\,)$. The transformed effect is simply $\hat{\kappa}_{IG} = T_{IG}/\sqrt{N}$, and the $100(1-\alpha)\%$ confidence interval for κ_{IG} is $\hat{\kappa}_{IG} \pm z_{1-\alpha/2}/\sqrt{N}$. Intervals of the same confidence for δ_{IG} are obtained by applying the back-trasnformation to the interval for κ_{IG}, namely $\delta_{IG} = \sqrt{2N}\,\sinh(\kappa_{IG}/\sqrt{2}\,)/\sqrt{n_T n_C}$.

To illustrate the meta-analysis of effect sizes we selected the data from 6 rows of Table 3 of Morris and DeShon (2002) which give the sample sizes and effect sizes $\delta_{IG,k}$, $k = 1, \ldots, 6$; see Table 2. Note that Study 4 contains slightly more evidence than Study 5 for a positive effect size, even though Study 4 is based on only 18 observations, compared to 395 observations for Study 5. This occurs because the estimated effect size in Study 4 is relatively large, 1.34, compared to that of Study 5, for which it is only 0.26. It is easy and meaningful to make such comparisons on the evidence scale of the T's, because all have the same standard normal error 1.

3.1 Meta-analysis for the FEM

In the remainder of Section 3 we drop the cumbersome subscripts $_{IG}$. We are interested in testing for heterogeneity of the effect sizes ($\hat{\delta}_k$'s) from Table 2. This is tested indirectly by testing heterogeneity of the transformed effect sizes ($\hat{\kappa}_k$'s), using the monotonicity of their relationship:

$$\kappa = \mathcal{K}\big(\sqrt{q(1-q)}\,\delta\big) = \sqrt{2}\,\sinh^{-1}\big(\sqrt{q(1-q)}\,\delta/\sqrt{2}\,\big)\ , \tag{3}$$

where $q = n_T/(n_T + n_C)$ and $\mathcal{K}$ is defined in (1).

For these $K = 6$ studies the weighted mean of the $\hat{\kappa}_k$'s in Table 2 is $\hat{\kappa}_w = 0.269$ and the value of Cochran's $Q^* = 19.5$. This exceeds the 0.95 quantile of the null distribution $\chi^2_{5,0.95} = 11.07$, so the hypothesis of homogeneity of transformed effect sizes is rejected at level 0.05, and one would usually adopt the REM.

However, for the sake of completeness of presentation, we first carry out the meta-analysis for the FEM. It is exactly the same as that for the paired-comparisons in Section 2, except that when transforming results back to the effect size scale we need to use the inverse of (3). For example, a point estimate of the common effect size δ, assuming $q = 0.5$, is $\hat{\delta} = 2\sqrt{2}\,\sinh(\hat{\kappa}_w/\sqrt{2}\,) = 0.541$. The 95% confidence interval for the common value of κ is $\hat{\kappa}_w \pm 1.96/\sqrt{N}$, or $[0.200, 0.339]$. After back-transformation to the effect size scale, the 95% interval for a common δ is $[0.400, 0.684]$.

3.2 Meta-analysis for the REM

For the REM, the DerSimonian and Laird estimate of the variance component γ^2 is $\hat{\gamma}^2_{DL} = 0.027$, which has square root $\hat{\gamma}_{DL} = 0.165$. The standard deviation of the reciprocals of the samples sizes n_k is $s_{1/n} = 0.019 < 0.165$, so we can be reasonably certain that the following results which are based on the Student-t statistic are reliable; see Chapter 25 of Kulinskaya, Morgenthaler and Staudte (2008).
The 95% confidence interval for κ is $[0.183, 0.556]$. We want an interval for a representative effect size $\delta = \sqrt{2}\,\sinh(\kappa/\sqrt{2}\,)/\sqrt{q(1-q)}$, which is just the inverse to (3). For this to be well defined we need to choose a representative q for the 6 studies. Since all the q_k are near 0.5, we take $q = 0.5$. Now a representative δ is well-defined, and so is the back-transformation to the effect size scale. The 95% confidence interval for δ is then $[0.367, 1.162]$.

4 Combining results from different designs

Morris and DeShon (2002) show that under certain conditions $\sqrt{2(1-\rho)}\,\delta_{RM} = \delta_{IG}$, and hence if one has an estimate of ρ, assumed to be constant for all studies, one can convert a repeated measures study estimate $\hat{\delta}_{RM}$ into an inter-group estimate $\hat{\delta}_{IG}$. This allows for the combination of results from both types of studies into one effect size point estimate. Confidence interval estimates are also possible, but further research is required to determine their accuracy, after allowing for the estimation of ρ.

References

Azorin, P.F. (1953). Sobre la Distribucion t no central I,II. *Trabajos de Estadistica*, **4** 173-198 and 307-337.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, **7** 177-188.

Kulinskaya, E., Morgenthaler, S. and Staudte, R.G. (2008). *Meta Analysis: a Guide to Calibrating and Combining Statistical Evidence* Chichester: John Wiley & Sons.

Morris, S.B. and DeShon, R.P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent group designs. *Psychological Methods*, **7**, 105-125.

Smoothing methods for analyzing spatial variability in heavy metal deposition

Silvia Suárez-Crespo[1], Rosa M. Crujeiras[1], Wenceslao González-Manteiga[1]

[1] Dpt. of Statistics and Operations Research, University of Santiago de Compostela, Spain

Abstract: Terrestrial mosses have been used as biomonitors for determining levels of heavy metal concentrations in the air. Concentrations from different heavy metals have been measured in a sampling network over Galicia (NW-Spain) in March and September 2006. The goal of this work is to describe the spatial variability of the concentrations and to test for differences between the two samples.

Keywords: heavy metal; smoothing methods; spatial trend.

1 Introduction

The accumulation of heavy metals over large areas and during long time periods may cause damage to living organisms. European and national regulations *Heavy Metals in European Mosses* project [Buse et al. (2003)] oblige to keep concentrations of heavy metals thoroughly controlled. In the region of Galicia (NW-Spain), heavy metal concentrations have been measured using mosses as biomonitors [Fernández et al. (2005)].
The moss technique for large-scale monitoring of long-range transport processes has been used for several decades as a means of surveying atmospheric heavy metal deposition, since the uptake of nutrients in mosses comes mainly from the air. A grid of 148 sampling points (see Figure 1), spaced by 15×15 kilometers, has been located over the region and limiting area. Concentrations of different metals have been measured in March and September 2006, and we will restrict our attention to Cobalt and Cadmiun concentrations, both measured in parts per billion (ppb).
The goal of this work is twofold: describe the spatial variability of the data and test if there exists any difference between the samples in spring (March) and after summer (September). In order to describe the spatial variability, measurements of Cobalt and Cadmiun are considered as realizations of spatial processes. Samples of Cobalt and Cadmiun in March and September have been previously normalized by a Box-Cox transformation, so the underlying spatial processes can be assumed to be Gaussian. The modelization procedure used is described as follows.
Consider a Gaussian process $\{Z(s),\ s \in D\}$, where $D \subset R^d$ and $s = (s_x, s_y)$ denotes a spatial location. The spatial variability of the process can be

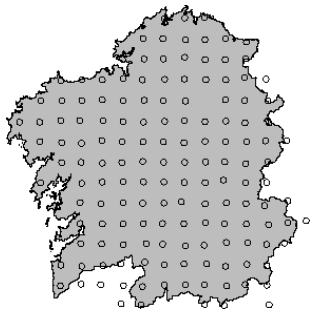


FIGURE 1. Sampling locations in Galicia and limiting area for the heavy metal concentrations.

represented by two different components: the large-scale variability (spatial trend) and the small-scale variability (spatial dependence), although the weight of each component cannot be derived from a single dataset. Hence, the spatial process can be written as:

$$Z(s) = m(s) + \epsilon(s), \quad s \in D$$

where $m(\cdot)$ denotes the spatial trend and ϵ is a zero-mean second-order stationary error process [Cressie (1993)]. Supose $\{Z(s_1), \ldots, Z(x_n)\}$ to be the observed data and $\omega_h(\cdot)$ the kernel function. The first step is to construct a nonparametric estimator of the trend, denoted by $\hat{m}(\cdot)$. For that purpose, we will consider the local linear estimator, which involves solving the problem:

$$\min_{\alpha,\beta} \sum_{i=1}^{n} (Z(s_i) - \alpha - \beta^T (s_i - s))^2 \omega_h(s_i - s)$$

and taking as the estimate $\hat{m}(s)$ the value of $\hat{\alpha}$ minimizing the previous expression [Bowman-Azzalini (1997)]. Once the trend has been estimated, residuals are obtained as $e(s) = Z(s) - \hat{m}(s)$. A test for independence [Diblasi (2001)] is applied to these residuals to test if there exists any variability which is not captured by the trend component. In case the variability is not fully described by the trend, this could be estimated from the residuals and included in a next step for re-estimating the trend.
For the practical implementation of the procedure, functions for nonpametric regression and testing for independence in spatial data in the R package `sm` have been used. In the next section, results obtained for Cadmium and Cobalt are shown.

2 Some results: Cobalt (Co) and Cadmium (Cd) analysis

After normalizing the measurements of Co and Cd, initial one-dimensional regression analysis for each metal over latitude and longitude have been carried out using the `sm.regression` function. The nonparametric regression curves for the two samples are compared. For this test both months are considered as groups and the observations are made at the same covariates values for each of them. Then, response values are computed and compared via a suitable statistic [Bowman-Azzalini (1997)]. Equality cannot be accepted for both Cd and Co samples (corresponding $p-$values smaller than 0.05, as a result of `sm.ancova`). For both metals, it can be obseved that the shape of the curves over latitude and longitude is similar for both months, but with a visible shift (higher values in September). For these cases (Co and Cd over latitude and longitude) parallelism is accepted.
In Figure 2, nonparametric surfaces for Cd in March and September are plotted. These surfaces are obtained using a cross-validatory bandwidth. Similar results are also given in Figure 3 for the Co. In addition, sequences of bandwidths have been tried for both cases and results remain similar. Residuals from the nonparametric estimated trends are tested for independence using `sm.variogram`. Independence is accepted applying the test introduced by [Diblasi-Bowman (2001)] in all the cases except for Co in September.
Finally, we will test if the surfaces for Cd and Co in the two months are equal. Again, equality is not accepted for both Cd and Co measurements. In these cases, as it was suggested by the one-dimensional analysis, a shift in the surfaces is clearly visible for Cd, and the parallelism of the trends is accepted. However, we reject parallelism for Co trends.
Higher concentrations of Cd and Co are found in September. This effect could be explained by the fact that not all the nutrients taken up by mosses remain in the organism. Actually, rain is quite frequent in Galician winter and spring, so a part of the metals absorbed by the mossess may be washed out by rain water.

Acknowledgments: This work has been supported by the Spanish Ministry of Science project MTM2008-03010 and FPI Grant BES-2009-025805. The authors want also to thank Jesús Aboal and José Ángel Fernández, from the Dpt. of Ecology and Cellular Biology in the University of Santiago de Compostela, who kindly provided the data.

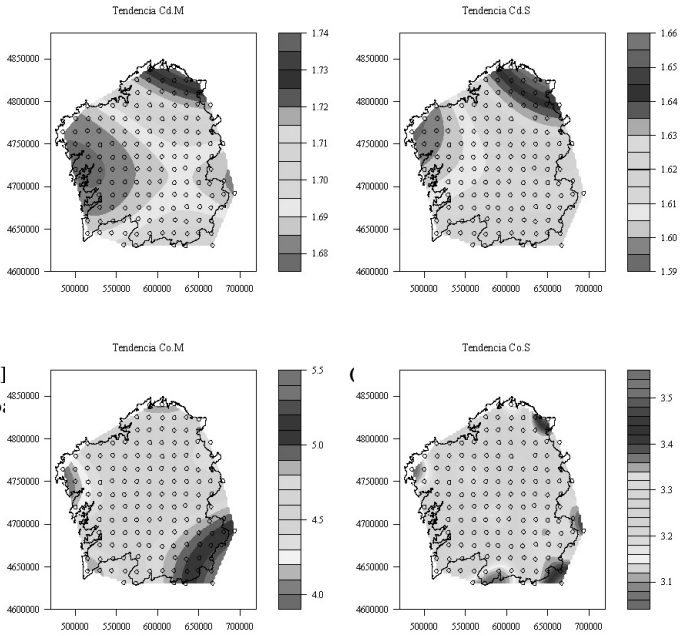


FIGURE 3. Nonparametric surfaces for Co in March (left panel) and September (right panel).

References

Buse, A., Norris, D., Harmens, H., Bker, P., Ashenden, T. and Mill, G. (2003). *Heavy metals in European mosses: 2000/2001 survey*. UNECE ICP Vegetation. United Kingdom: Bangor.

Cressie, Noel A. C. (1993). *Statistics for spatial data*. New York: John Wiley & Sons Inc.

Fernández, J.A., Real, C., Couto, J.A, Aboal, J.R. and Carballeira, A. (2005). The effect of sampling design on extensive bryomonitoring surveys of air pollution. *Science of the Total Environment*, **337**, 11-21.

Diblasi, A. and Bowman, A. W. (2001). On the use of the variogram in checking for independence in spatial data. *Biometrics*, **57**, 211–218.

Strategies for local smoothing in high dimensions: using density thresholds and adapted GCV

James Taylor[1], Jochen Einbeck[1]

[1] Department of Mathematical Sciences, University of Durham, Durham, DH1 3LE, UK, {james.taylor1, jochen.einbeck}@durham.ac.uk

Abstract: Local polynomial fitting for univariate data has been widely studied and discussed, but up until now the multivariate equivalent has often been deemed impractical, due to the so-called *curse of dimensionality.* Here, rather than discounting it completely, we use density as a threshold to determine where over a data range reliable multivariate smoothing is possible, whilst accepting that in large areas it is not. An adapted version of generalized cross-validation for multivariate bandwidth selection is also discussed.

Keywords: Smoothing, Density, Threshold, Bandwidth

1 Introduction

We are given d-dimensional covariates $X_i = (X_{i1}, ..., X_{id})^T$ and response values Y_i where $i = 1, ..., n$. Local polynomial regression is a nonparametric way of estimating the mean function $m(x) = E(Y|X = x)$. Assumed is that

$$Y_i = m(X_i) + \epsilon_i \tag{1}$$

where ϵ_i are random variables with zero mean and variance σ^2. We concentrate on local linear regression where hyperplanes of the form $\beta_0 + {\beta_1}^T x$, where β_1 and x are both vectors, are fitted locally. For each point x in d-dimensional space one minimizes

$$\sum_{i=1}^{n} \left\{ Y_i - \beta_0 - \sum_{j=1}^{d} \beta_{1j}(X_{ij} - x_j) \right\}^2 K_H(X_i - x) \tag{2}$$

with respect to $\beta = (\beta_0, \beta_{11}, ..., \beta_{1d})^T$, yielding the estimator of the mean function $\hat{m}(x) = \hat{\beta}_0$. Here, $K_H(x) = |H|^{-1/2} K(H^{-1/2}x)$ where K is a multivariate kernel function and H is the bandwidth matrix. For K, we use primarily a product of Gaussian kernels since this is the least temperamental kernel function in regions where data is sparse, which occur more often in higher dimensions. H is crucial in determining the amount and direction of

smoothing, and we choose to use a diagonal matrix, $H = \text{diag}(h_1^2, ..., h_d^2)$, for computational ease. We have adjusted generalized cross-validation slightly for use with multivariate data, and this is detailed in Section 3.
The problem primarily addressed here is the *curse of dimensionality* which refers to the issues that arise when data becomes very sparse in higher dimensions. If there is not sufficient data in a neighbourhood, then the variance of the fit is too high, or with some kernel functions, such as the popular Epanechnikov kernel, the calculations just break down completely. Often, local polynomial fitting is abandoned as a result of these problems, and other methods such as the additive models suggested in Hastie and Tibshirani (1990), are favoured. Local polynomial fitting however has the big advantage of being considerably more flexible. In Section 2, we pursue a technique to avoid the curse of dimensionality, enabling us to achieve the best possible estimate of m where sufficient information is available.

2 Density as a threshold

The method is one which essentially ignores all neighbourhoods which don't contain enough data, and so only performs smoothing over some region in which estimation is considered reliable, where the bias and variance of $\hat{m}$ can be kept reasonably low. In this way the curse of dimensionality is avoided. This method is not universal in the sense that it doesn't give estimates over the whole data range, but it is satisfactory in the sense that it gives estimates, with all the advantages of local polynomial regression, in some areas. To find these areas, and to discover where there is enough data, we examine the density f of X. The density estimate for a multivariate point x is;

$$\hat{f}(x) = n^{-1} \sum_{i=1}^{n} K_H(x - X_i) \tag{3}$$

In calculating the density, again a bandwidth matrix is needed, and for our purposes it is advisable to use the same parameters here as in the regression, for reasons which will become clear. We seek a threshold T such that, if at point x we have $f(x)>T$, then an estimate using local linear regression can be considered somewhat reliable, and otherwise, care should be taken and an alternative method sought, possibly local constant fitting. According to Loader (1999), one has $\frac{1}{\sigma^2}\text{Var}(\hat{m}(X_i)) \leq \text{infl}(X_i) \leq 1$. Hence, bounding the influence implies bounding the variance. Using the asymptotic approximation of the influence function given in Loader (1999), a natural choice of T is straightforwardly derived from the latter inequality;

$$T = \frac{\rho K(\mathbf{0})}{n \prod_{i=1}^{d} h_i} \tag{4}$$

where

$$\rho = e_1^T \left(\int_a^\infty K(\mathbf{v}) A(\mathbf{v}) A(\mathbf{v})^T d\mathbf{v} \right) e_1, \tag{5}$$

$\mathbf{v} = (v_1, ..., v_d)^T$ and $A(\mathbf{v}) = (1, \mathbf{v})^T$. The position of the bandwidth parameters in the denominator is justified since with larger bandwidth parameters, in areas of relative high density, the density at x will be lower than with smaller parameters, and so a lower threshold is needed.
The parameter a appearing in the lower integral limit reflects the distance to the boundary of f for which the criterion is optimized. Based on extensive testing in the local linear case we recommend the value $a = -0.85$, corresponding to a point situated $0.85h_i$ inside the boundary. This is quite intuitive as this is just about the region where one would assume data sparsity to become a problem.

3 Adapted GCV

The *curse of dimensionality* causes problems in the area of bandwidth matrix selection too. We believe in the use of a classical method over a plug-in one due to the reliance on asymptotics of the latter. The asymptotic assumption of bandwidths tending to zero seems to be inappropriate in order to select the relatively large bandwidths needed for multivariate local smoothing.
One such classical method is generalized cross-validation which is less precise than other cross-validation, but computationally less demanding. We propose an adaptation to this which, using the median and weighting, removes the influence of data points in less dense areas which otherwise may have a disproportionate effect on the procedure, and can cause extreme values of h_i being chosen. This effect is more likely to occur as d increases. The minimization of the below has been trialled with some success;

$$AGCV(H) = n^{-1} \sum_{i=1}^{n} \left\{ \frac{Y_i - \hat{m}_H(X_i)}{1 - \psi} \right\}^2 w(X_i) \tag{6}$$

where ψ is the median of the diagonal elements of the smoother matrix after excluding the elements contributed by the points for which $w(X_i) = 0$. We set $w(X_i) = 1$ for all i except the r points at which $f(X_i)$ are smallest, at which it is 0. r is the number of points which could be considered isolated i.e. where the density at that point is equal to the density of just one data point. This is best examined using Epanechnikov kernels. The bandwidth parameters to be used in the density estimation here should be the optimal values calculated from an external source such as the *np* package in R. Choosing $r > 0$ is both a matter of finetuning by focussing on the denser region in which we are interested, and also removing any computational constraint imposed by points in sparser regions.

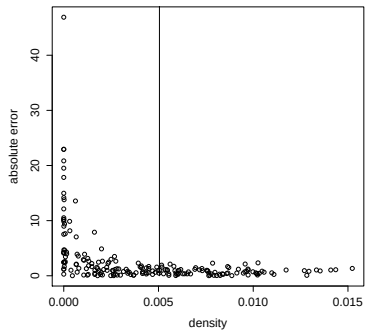


FIGURE 1. $|m(X_i) - \hat{m}(X_i)|$ v. $\hat{f}(X_i)$

4 Simulation

We simulated 3-dimensional covariates through a t-distribution with 2 degrees of freedom centered at 15.5. The response values were generated according to the model (1) with $m(X_i) = -12\cos(X_{i1}) + 5\sin(5X_{i2}) + 10\log(X_{i3}) + 17$ and $\epsilon_i \sim N(0,1), i = 1, ..., 500$. 300 of these points were used to estimate m while the remaining 200 were used to test the threshold method. In this simulation, of the 200 points tested, the threshold of 0.00505 excluded 121 of them. The value of $a = -0.85$ consistently excluded the poorest performing points, whilst deeming the better points, in terms of the absolute error, $|m(x_i) - \hat{m}(x_i)|$, apt for smoothing. This is shown in Fig.1. AGCV was also successful with this simulation. The usual GCV method suggested some h_i greater than the data range which is clearly unacceptable, but with $r = 39$, AGCV suggested a much more reasonable $H = \text{diag}(0.347^2, 0.129^2, 2.92^2)$. The density bandwidth parameters were selected using the *np* package. Simulations were performed satisfactorily with 1,3 and 16-dimensional covariates. The values of ρ calculated for use in the simulations were 1.5, 3.12 and 147.3 respectively.

References

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

Loader, C. (1999). *Local Regression and Likelihood.* Springer.

Accounting for a non-ignorable tracing mechanism in a retrospective breast cancer cohort study

Andrew C. Titman[1], Gillian A. Lancaster[1], Katie Carmichael[2], Diane Scutt[2]

[1] Department of Mathematics and Statistics, Lancaster University, Bailrigg, LA1 4YF, United Kingdom.
[2] School of Health Sciences, Liverpool University, Liverpool, L69 3GB, United Kingdom.

Abstract: Motivated by a retrospective breast cancer cohort study, we consider the analysis of competing risks data where tracing of patients is dependent on survival to some truncation time. By modelling post-cancer survival within a multi-state model framework, the tracing bias can be accounted for. Two approaches are considered, a likelihood based method using piecewise-constant transition intensities under a Markov assumption and a pseudo-likelihood method using inverse probability of tracing weights. For the breast cancer example, the methods improve the precision of estimates compared to a conventional approach based on excluding patients.

Keywords: non-ignorable tracing; multi-state model; pseudo-likelihood; breast cancer.

1 Introduction

The Merseyside Breast cancer study involves 12942 women self-referred to the Liverpool Breast Screening Unit between 1979 and 1987. The aim of the original study was to establish whether mammographic density (parenchymal patterns defined using Wolfe's criteria (Wolfe, 1976)) were associated with hypothesised risk factors for breast cancer. The women completed a detailed questionnaire to measure known risk factors and in addition undertook a mammogram.

A case-control study which used the Merseyside and Cheshire Cancer Registry to obtain breast cancer cases from the original cohort and aged matched controls randomly taken from the cohort, identified an apparent association between breast volume asymmetry and breast cancer (Scutt et al, 2006). In 2006, to validate the results of the case-control study, the women were traced through the UK Office for National Statistics, using the Central Health Register Inquiry System (CHRIS). Tracing based on CHRIS

is considerably cheaper than a full manual trace due to being fully automated. However, CHRIS was established in 1991, and generally only included women currently registered with a GP at that time. Therefore 1991 is a truncation time, patients who died or emigrated before this time would not be on the system. For traced patients we have data on times of cancer diagnosis before and after 1991.
We take a multi-state modelling approach (Andersen and Keiding, 2002) to account for the tracing mechanism. A four state process, $X(t)$, is used with states representing no disease, breast cancer, other cancer and death. The requirement of survival to a particular time leads to a *purged process* (Hoem, 1969). This theory can be used to characterise the bias caused by not accounting for the tracing mechanism.

2 Methods

2.1 Standard Approach

To avoid problems of tracing we can restrict ourselves to a subset of the dataset, though this can lead to a significant loss of information. We only include traced patients who are healthy in 1991. These remaining patients then form a standard, left-truncated, cohort. Standard competing risks methods can be used based on modelling of the cause-specific hazard functions (Putter et al, 2007).

2.2 Full likelihood

If the only cause of patients being untraced was through lack of survival to time t_{iu}, then the untraced patients would effectively have interval-censored death times between $[t_{il}, t_{iu}]$. However, in practical applications lack of tracing may be due to additional non-informative factors. As a result we cannot infer that untraced individuals died before t_{iu} and therefore provide no information.
The full likelihood can be written as

$$L(\theta) = \prod_{i=1}^{\tilde{N}} \frac{L_i(\theta)}{1 - p_{0R}(t_{il}, t_{iu}; \theta)}$$

where $L_i(\theta)$ is the standard likelihood contribution for individual i from data without tracing error, t_{il} and t_{iu} are the entry time and truncation time for individual i, $p_{0R}(t, s) = P(X(s) = R | X(t) = 0)$, $\tilde{N}$ is the number of traced patients, 0 represents the disease free state and R is death. In most cases the full likelihood is intractable. However, a tractable weakly parametric form is to specify the baseline intensities as piecewise constant between pre-specified change points and make a Markov assumption regarding the transition intensities $\lambda_{rR}(t)$, $r = 1, \ldots, R-1$. Calculation of

$p_{0R}(t,s)$ is then reasonable provided the number of change points is not too large. Covariates can be included by assuming a proportional intensities model.

2.3 Pseudo-likelihood

To avoid the problems associated with computation of the full likelihood, particularly in the semi-parametric case, a pseudo-likelihood approach can be adopted (Kalbfleisch and Lawless, 1988). Under a proportional intensities assumption with non-parametric baseline hazard, a corresponding pseudo-partial likelihood can be derived (Johansen, 1983) of the form

$$pl_p(\beta) = \sum_{i=1}^{N} \int_0^{\tau} \Delta_i p_i^{-1} \left(\exp(\beta^T z_i) - \log(S_p^{(0)}(\beta, t)) \right) dN_i(t)$$

where Δ_i is a sampling indicator, N is the total population size, $S_p^{(0)}(\beta,t) = \sum_{i=1}^{N} \Delta_i p_i^{-1} Y_i(t) \exp(\beta^T z_i)$, $Y_i(t)$ takes value 1 if subject i is at risk at time t and is zero otherwise and $N_i(t) = I(T_i \leq t)$.
In the context of data subject to informative tracing bias, the sampling probabilities p_i refer to tracing probabilities. The tracing probabilities are calculated by generalising the approach of Copas and Farewell (2001) to a competing risks setting and involve modelling the post-illness survival and study entry time distributions.
In the competing risks model the possible first events are censored from healthy, death from healthy, development of illness r for $r = 1, \ldots, R-1$. If a subject is censored or dies from healthy then their tracing probability is

$$p_i(T, \delta \in \{0, R\}) = P(T + X \geq t_{iu}) = 1 - F_X(t_{iu} - T).$$

This reflects that the subject will only have been traced if the first event occurs after the truncation time t_{iu}. If a subject's first event is development of risk r at time T then

$$p_i(T, \delta = r) = P(T + X \geq t_{iu}) + \int_0^{t_u - T} S_{rR}(x + T, t_{iu}) G(x + T, t_{iu}) dF_X(x)$$

where $S_r(u, t_{iu})$ denotes the survival distribution from risk r to time t_{iu} given entry at time u, while $G(u, t_{iu})$ denotes the censoring survival distribution. This reflects that a patient who develops the illness before t_{iu} must survive until t_{iu} and also not be censored.

3 Application to Breast Cancer Cohort Study

For the Merseyside Breast Cancer study, we are interested in the covariates affecting the rate of onset to breast cancer. We consider only times to first

TABLE 1. Estimated covariate effects on log-linear scale and associated standard errors using the three methods of estimation.

	Post-1991		Full likelihood		Pseudo-likelihood	
Covariate	Estimate	SE	Estimate	SE	Estimate	SE
Vol. asymmetry	-0.035	0.101	-0.038	0.076	-0.035	0.090
Par. type:N1	0		0		0	
Par. type:P1	0.669	0.288	0.425	0.229	0.389	0.239
Par. type:P2	0.981	0.287	0.822	0.228	0.792	0.237
Par. type:DY	1.005	0.269	0.836	0.210	0.803	0.219
Family history	0.319	0.145	0.399	0.125	0.409	0.127
Previous biopsy	0.273	0.204	0.261	0.176	0.258	0.183
Age at menarche	-0.021	0.046	-0.066	0.035	-0.070	0.040
Height (per 10cm)	0.137	0.106	0.111	0.094	0.112	0.097

cancer, and we treat other cancers as a single competing event to breast cancer. In addition, all women are at risk of death. Age is taken as the primary time scale. In all three analyses we assume covariates affect the transition intensity from healthy to breast cancer based on a proportional intensities assumption.

For the standard analysis, 268 women are excluded including 72 cases of breast cancer (out of a total of 341) because of events before 1991. For the pseudo-likelihood method, the hazard of death from cancer type $r = 1, 2$ at time t was taken as $\lambda_{rR}(t, t^*) = \lambda_{rR}(t - t^*) \exp(\beta_r t^*)$ where t^* is the time of cancer onset. This represents a semi-Markov assumption.

Table 1 shows that the estimated covariate effects are broadly similar between the three methods. The standard analysis based on women healthy in 1991 gives slightly higher estimates of the effect of denser parenchymal pattern types (P1, P2 and DY). There is a general improvement in the precision of the estimates in the methods using the pre-truncation data. For the full likelihood method the standard errors are around 15-25% lower than through the standard competing risks analysis. For the pseudo-likelihood method the estimated standard errors are slightly higher than those of the full likelihood approach, but still give around a 10-20% improvement compared to the standard competing risks analysis. Standard errors for the post-1991 and pseudo-likelihood methods were based on 1000 bootstrap samples. For the full likelihood approach standard errors were based on inverting the Fisher information.

4 Conclusion

Modelling of time to death from illness can allow more efficient estimation in cohort studies subject to survival related tracing bias. A para-

metric full-likelihood approach is possible as well as a semi-parametric pseudo-likelihood method. In the breast cancer cohort study considered, both methods gave similar results, including in terms of the reduction in standard errors.

A disadvantage of the pseudo-likelihood approach is that it requires modelling of the entry times into the study. This was not problematic for the breast cancer cohort study because the mammogram dates of all women originally recruited were known and hence the distribution could be estimated empirically. More generally, this may not be the case. Moreover, the pseudo-likelihood method is not appropriate for a study where all subjects are recruited at the same date. The full likelihood method does not require modelling of entry times, but does rely on parametric and Markov assumptions.

Acknowledgments: This work was supported by a grant from Cancer Research UK (C24779/A9646).

References

Andersen P.K., and Keiding N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, **11**, 91-115.

Copas A.J., and Farewell V.T. (2001). Incorporating retrospective data into an analysis of time to illness. *Biostatistics*, **2**, 1-12.

Hoem J.M. (1969). Purged and partial Markov chains. *Skandinavisk Aktvarietidsskrift*, **52**, 147-155.

Johansen S. (1983). An extension of Cox's regression model. *International Statistical Review*, **51**, 165-174.

Kalbfleisch J.D., and Lawless J.F. (1988). Likelihood analysis of multi-state models for disease incidence and mortality. *Statistics in Medicine*, **7**, 149-160.

Putter H., Fiocco M., and Geskus R.B. (2007). Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine*, **26**, 2389-2430.

Scutt D., Lancaster G.A., and Manning J.T. (2006). Breast asymmetry and predisposition to breast cancer. *Breast Cancer Research*, 2006.

Wolfe JN. (1976). Breast patterns as an index of risk for developing breast cancer. *American Journal of Roentgenology*, **126**, 1130-1139.

Estimating food expenditure in small areas using the Spanish Household Budget Survey

M. D. Ugarte[1], T. Goicoa[1], A. F. Militino[1]

[1] Departamento de Estadística e I.O., Universidad Pública de Navarra

Abstract: Small area estimators based on a penalized spline regression model are obtained. In each small area, individual curves are fitted using penalized splines with B-spline bases. The prediction error of the proposed estimators is also derived. To account for possible bias, a bootstrap correction is obtained.The methods are used to estimate the percentage of food expenditure for alternative household sizes at provincial level using the 2006 Spanish Household Budget Survey.

Keywords: P-splines; Model-based estimators; Prediction error; Bias correction

1 Introduction

Small area estimation techniques have experienced a quick evolution in the last few years motivated by the necessity of precise information of different nature for small domains. Likely, the most important area to apply and develop small area estimation tools is official statistics. Statistical institutes are in charge of conducting the most important surveys and reporting estimates about economic, demographic, agricultural, and many other issues related to the state of a country. However, these surveys have been traditionally designed to provide reliable estimates for large areas or domains of interest. Consequently, to derive reliable information at smaller levels one must develop small area estimation techniques.

The key point in small area estimation is the use of models to provide reliable estimates for small areas. In particular, P-spline models are a powerful tool for modelling non-linear but smooth relationships between variables. In this work, we construct a small area estimator based on a P-spline model that uses B-spline basis (see Eilers and Marx, 1996). The model considers a different spline for each small area. We take advantage of the mixed model representation of the P-splines to express the model as a linear component common to all small areas (the fixed part of the mixed model), and a non linear component (the random part of the mixed model) specific for each small area. Thus, the non linear part of the spline can be seen as a specific small area effect. The resulting model leads to a block-diagonal covariance structure, making the computations easier. The procedure is used to estimate the percentage of food expenditure relative to total expenditure in

208 small areas in Spain, using data from the 2006 Spanish Household Budget Survey (SHBS). This survey provides annual information on the nature and destination of consumption expenditures, as well as on a range of features related to household living conditions. One of the main objectives of the SHBS is estimating the aggregate annual consumption expenditure of households for the whole of Spain and its Autonomous Regions, as well as their classification by different household variables.

1.1 Small area estimator

Suppose we have I small areas and n_i observations within each area. Consider the following model

$$y_{ij} = f_i(x_{ij}) + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma_e^2 x_{ij}^{2\delta}), \quad i = 1, \ldots, I, \quad j = 1, \ldots, n_i.$$

where f_i is a smooth function for each area which can be approximated sufficiently well using P-splines, and δ is a parameter to account for possible heteroscedasticity. Using the mixed model representation of P-splines the proposed model can be expressed as

$$y_{ij} = f_i(x_{ij}) + \epsilon_{ij} \cong \beta_0 + \beta_1 x_{ij} + \sum_{j=1}^{J} u_{ij} Z_{ikj} + \epsilon_{ij}, \tag{1}$$

where u_{ij} is the jth random effect corresponding to the spline for the ith small area, and Z_{ikj} is the kjth element of the matrix of random effects obtained from the mixed model representation of the spline for the ith area. In matrix form, Model 1 is expressed as

$$Y = X\beta + Zu + \epsilon, \quad u \sim N(0, \Sigma_u = \sigma_u^2 I_{L\times L}), \quad \epsilon \sim N(0, \Sigma_e = \sigma_e^2 R) \tag{2}$$

The quantity of interest is the small area mean

$$\bar{Y}_i = \bar{X}_i \beta + \bar{Z}_i u_i.$$

where $\bar{X}_i = (1, \bar{X}_i)$, $\bar{X}_i$ is the known population mean of the explanatory variable for the small area i, and $\bar{Z}_i$ is the mean of the elements corresponding to area i of the matrix Z derived from a B-spline basis constructed for the whole population.

1.2 MSE and bias correction

As the P-spline model leads to a block diagonal covariance matrix, results provided by Das *et al.* (2004) can be applied. However, as the P-spline approach can lead to a biased estimation of the true functional form of

the relationship between the response and the covariate, a bias correction is proposed. As the form of this bias is unknown, it is difficult to model it. However, if the bias is suspected to be high, it is necessary to correct it, as point estimates are important quantities for statistical institutes. For example, in the analysis of the 2006 Spanish Household Budget Survey that will be done in the next section, good point estimates are important as they will be used for the National Accounts. In this work, a bootstrap correction of bias is considered.

2 Application

In this section, data from the 2006 Spanish Household Budget Survey (SHBS) are analyzed. The interest relies on estimating the percentage of food expenditure relative to the total expenditure (in 10000 euros) for the 52 Spanish provinces and different household sizes: households with one person, two people, three people, and four or more people. The small areas are defined then by the combination of provinces and household sizes giving a total of $52 \times 4 = 208$ small areas. Traditionally, linear mixed effects models have been considered in small area estimation, but a look to the scatter plots (not shown here) of the percentage of food expenditure against total expenditure for each small area reveals that the relationship is not linear as the percentage of food expenditure seems to decline for higher levels of total expenditure. Hence, a small area estimator based on P-spline model described before is derived. We also construct an estimator based on a model that considers a single curve in all the areas but we finally discard it as the behavior of many of the small areas were different.
We have obtained good estimates and coefficients of variation for all small areas. There are two provinces with a higher percentage of food expenditure: Ourense and Cádiz. This result matches with other economical indicators as they are quite deprived provinces where the percentage of food expenditure is important. In contrast, recognized richer regions, such as Madrid, Navarra, País Vasco, and Cataluña display lower percentages of food expenditure.

Acknowledgments: This work has been supported by the Spanish Ministry of Science and Innovation (MTM 2008-03085)

References

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

Das, K., Jiang, J. and Rao, J. N. K. (2004). Mean squared error of empirical predictor. *Annals of Statistics*, **32**, 818-840.

Model Selection for Clustered Data: Conditional AIC under Generalized Linear and Proportional Hazards Mixed Models

Florin Vaida[1], Michael Donohue[2], Ronghui Xu[2], Rosanna Haut[2]

[1] communicating author; for oral presentation,
[2] Division of Biostatistics and Bioinformatics (FV, MD, RX) and Department of Mathematics (RX, RH), University of California at San Diego, La Jolla, CA 92093, USA; fvaida@ucsd.edu, mdonohue@ucsd.edu, rxu@math.ucsd.edu, rhaut@math.ucsd.edu

Abstract: This paper focuses on the issue of model selection for generalized linear mixed models (GLMM) and proportional hazards mixed models (PHMM). For linear mixed models (LMM) used with clustered data, Vaida and Blanchard (2005) proposed the conditional Akaike information (cAI) and the corresponding criterion (cAIC), when the focus is on the clusters. Here we show how to extend the approach to GLMM and PHMM. While the derivation for the LMM used exact calculation, in the more general case asymptotics and approximations are used. We apply the methods to the analysis of two cancer clinical trials.

1 Introduction

Mixed effects models have been widely used to analyze clustered data. Often, the focus of inference is on the random effects themselves. For example, in a multicenter clinical trial where the treatment effect is heterogeneous among centers, it is of scientific interest to estimate the treatment effects themselves and to investigate the cause of the treatment differences. Other examples of cluster level focus occur in ecology, small-area estimation, and animal husbandry.

Consider the independent n clusters scenario, with m_i observations in each cluster $i = 1 \dots n$. Conditionally on the cluster-specific random effects b_i, the outcomes y_{ij} are independent and follow a GLMM with mean

$$\mu_{ij} = E(y_{ij}|b_i) = g^{-1}(\beta' x_{ij} + b_i' z_{ij}), \tag{1}$$

where x_{ij} and z_{ij} are the covariate vectors for the fixed effects β and random effects b_i of cluster i, $b_i \sim N(0, \Sigma)$, and g is the link function. For cluster level inference, any future prediction takes place in the same clusters as the observed data, and the random effects for these clusters are held constant.

More specifically, let y^0 be independently replicated outcomes from the same conditional distribution as the original data y given the same random effects b. Here y, y^0 and b are random vectors consisting of elements y_{ij}, y^0_{ij} and b_i, respectively. Vaida and Blanchard (2005) defined the conditional Akaike information:

$$\text{cAI} = -2E_{(y,b)}E_{y^0|(y,b)}\{l(y^0|\hat{\beta}(y),\hat{b}(y))\}, \tag{2}$$

where $l(\cdot|\cdot)$ is the conditional log-likelihood under the model, and $\hat{\beta}(y)$, $\hat{b}(y)$ are estimators of β and b based on the data y, for example, the maximum likelihood and the empirical Bayes estimators. For LMM with known variance components cAI has the unbiased estimator

$$\text{cAIC} = -2l(y|\hat{\beta}(y),\hat{b}(y)) + 2\rho\ ; \tag{3}$$

where the bias correction factor ρ is the effective degrees of freedom of the LMM of Hodges and Sargent (2001) and Ye (1998). The cAIC is referred to as the *conditional Akaike Information criterion*. Vaida and Blanchard (2005) and Liang, Wu and Zou (2008) give formulas for ρ in the more general case of unknown variance parameters. They also discuss the connections of cAIC with the Deviance information criterion of Spiegelhalter *et al.* (2002). In this paper we develop cAIC under the GLMM and the proportional hazards mixed-effects model (PHMM). Exact calculation is not available outside normal linear models, and asymptotic approximations are necessary. An additional concern is the presence of nuisance parameters under the PHMM, in this case of infinite dimension.

2 Conditional AIC for Generalized Linear Mixed Models

In the GLMM given by (1), write $D = \text{var}(b) = \text{diag}_n(\Sigma)$, the block-diagonal matrix with n blocks equal to Σ. Let X_i and Z_i be the matrices with rows x'_{ij} and z'_{ij}, and let $X = (X'_1, \ldots, X'_n)'$ and $Z = \text{diag}(Z_1, \ldots, Z_n)$ be the $(N \times p)$ and $(N \times q)$ model matrices, where p is the length of β, $q = nd$, $N = m_1 + \ldots + m_n$. Further, let w_{ij} be the GLMM weights given by $w_{ij} = [\text{var}(y_{ij}|b_i)\{g'(\mu_{ij})\}^2]^{-1}$, and $W = \text{diag}(w_{ij})$. Then the following result holds:

Theorem 1 (cAIC for GLMM)
Under suitable regularity conditions, the conditional Akaike information criterion is given by

$$\text{cAIC}_1 = -2l(y|\hat{\beta},\hat{b}) + 2\rho_1, \qquad \text{where} \tag{4}$$

$$\rho_1 = \text{tr}\left\{\begin{pmatrix} X'WX & X'WZ \\ Z'WX & Z'WZ \end{pmatrix}\begin{pmatrix} X'WX & X'WZ \\ Z'WX & Z'WZ + D^{-1} \end{pmatrix}^{-1}\right\}. \tag{5}$$

As $n \to \infty$ and $\min_i m_i \to \infty$, cAIC_1 is consistent for cAI (2), in the sense that $E(\text{cAIC}_1) = \text{cAI} + E(R)$, where the remainder term $R = o_p(1)$. If R is uniformly integrable, then $E(\text{cAIC}_1) \to \text{cAI}$. In addition, $p \le \rho_1 \le p + q$. The bias correction ρ_1 can be also written as $\rho_1 = \text{tr } A$, with $A = \text{var}(\dot{l}(y,b)|b)\{E(\ddot{l}(y,b)|b)\}^{-1}$, where $\dot{l}$ and $\ddot{l}$ are the first and second derivatives with respect to β of the joint log-likelihood $l(y|\beta,b) + \log p(b|D)$.

3 cAIC for PHMM

The PHMM (frailty model) is given by $\lambda_{ij}(t) = \lambda_0(t)\exp(\beta' x_{ij} + b_i' z_{ij})$, where $\lambda_{ij}(t)$ is the hazard function of observation j from cluster i, $i = 1 \dots n$, and $b_i \sim N(0, \Sigma)$ as before; $\lambda_0(t)$ is a nuisance parameter. The outcome data are $y_{ij} = (Y_{ij}, \delta_{ij})$, where Y_{ij} is the right-censored failure time and δ_{ij} is the event indicator. For conditional focus, we can profile out the baseline hazard and arrive at the conditional partial log-likelihood

$$\text{pl}(y|\beta,b) = \sum_{i=1}^{n}\sum_{j=1}^{n_i} \delta_{ij} \log \frac{\exp\{\beta' x_{ij} + b_i' z_{ij}\}}{\sum_{i'j'} \exp\{\beta' x_{i'j'} + b_{i'}' z_{i'j'}\}}, \tag{6}$$

where the sum in the denominator is over all i', j' such that $Y_{i'j'} > Y_{ij}$. We define the conditional Akaike information under PHMM, similarly to (2) as

$$\text{cAI} = -2E_{f(y,b)}E_{f(y^0|b)}\{\text{pl}(y^0|\hat{\beta}(y), \hat{b}(y))\}. \tag{7}$$

A result equivalent to that established for GLMM holds:

Theorem 2 (cAIC for PHMM)
Under suitable regularity conditions,

$$\text{cAIC}_2 = -2pl(y|\hat{\beta}, \hat{b}) + 2\rho_2 \tag{8}$$

is a consistent estimator of the cAI (7) in the sense of Theorem 1, with $\rho_2 = \text{tr}\{\text{var}(\dot{pl}(y,b)|b)\{E(\ddot{pl}(y,b)|b)\}^{-1}\}$. A formula similar to (5) applies. In practice, ρ_2 can be estimated either directly from (8) or via resampling (bootstrap). Simulation studies (not included) show that both methods are robust and accurate. Similar estimators were derived, from different considerations, by Ha, Lee and MacKenzie (2007).

4 Two Clinical Trials for Cancer

The *Skin Cancer Prevention Study* was a randomized, double-blinded, placebo-controlled clinical trial of beta-carotene to prevent non-melanoma skin cancer in high risk subjects (Greenberg *et al.*, 1990). The 1805 subjects were randomized to beta-carotene or placebo and examined yearly

for five years. The outcome is the number of new skin cancers per year. We analyzed 1683 subjects with complete data. We considered six Poisson mixed-effects models (Table 1), including fixed effects for four covariates (treatment, age, gender, skin type), with or without a linear and quadratic time trend, random subject effect, with or without random subject-specific time effect.The quadratic effect of time is highly significant; cAIC_1 clearly chooses models with random intercept only. The number of degrees of freedom ρ_1 for the best model (7,1) is 1638, indicating strong heterogeneity among patients even after controlling for covariates.

TABLE 1. Skin cancer prevention study. Estimates, likelihood, and cAIC from six Poisson mixed-effects models. Standard errors of estimates are omitted.

	Model	(5,2)	(6,2)	(7,2)	(5,1)	(6,1)	(7,1)
Intercept	β	-4.86	-4.79	-5.36	-4.28	-4.35	-4.18
Age		0.02	0.02	0.02	0.02	0.02	0.02
Skin		0.33	0.33	0.32	0.33	0.33	0.33
Gender		0.69	0.69	0.70	0.63	0.63	0.63
Exposure		0.19	0.19	0.19	0.18	0.18	0.18
Year		–	-0.03	0.38	–	0.02	-0.13
Year^2		–	–	-0.08	–	–	0.03
Intercept	σ^2	10.30	9.97	13.12	2.37	2.38	2.39
Year		0.84	0.85	1.22	–	–	–
$-2l(y\|\hat{\beta},\hat{b})$		4942.93	4942.28	4825.32*	6072.10	6068.57	6063.50
cAIC_1		11345.03	11345.31	11283.33	9342.79	9341.87	9339.13*

E1582 Lung Cancer Trial. This multi-center lung cancer study enrolled 579 patients from 31 institutions, randomized to standard chemotherapy (CAV) or an experimental treatment (CAV-HEM). The primary endpoint was time to death. Covariates included presence of bone metastases, presence of liver metastases, performance status at entry and weight loss prior to entry. Gray (1995) and Vaida and Xu (2000) found a significant difference in treatment effect between the 31 institutions, treated as clusters. Here we consider three models , all of them including the five important covariates. Models (5,0), (5,1), (5,2) include no random effects, random treatment effect, and random treatment and bone metastases effects (assumed independent) (Table 2). cAIC_2 indicates a slight preference for the more parsimonious model (5,0).

References

Gray, R. (1995). Tests for variation over groups in survival data. *J Am Stat Assoc.* **90**, 198–203.

TABLE 2. E1582 lung cancer trial data. Estimates, standard errors, and conditional AIC for three models.

		(5,0)	(5,1)	(5,2)
Treatment	β	−0.254 (0.085)	−0.250 (0.104)	−0.247 (0.119)
Bone met.		0.223 (0.093)	0.212 (0.095)	0.230 (0.144)
Liver met.		0.429 (0.090)	0.423 (0.091)	0.393 (0.094)
Perf. status		−0.602 (0.104)	−0.641 (0.109)	−0.649 (0.131)
Wt. loss		0.200 (0.087)	0.218 (0.089)	0.208 (0.092)
Treatment	σ^2	-	0.071 (0.069)	0.046 (0.184)
Bone met.		-	-	0.129 (0.083)
cAIC_2		6106.50*	6110.06	6111.68

Greenberg, E. R., Baron, J. A., Stukel, T. A. *et al.* (1990) A clinical trial of beta carotene to prevent basal-cell and squamous-cell cancers of the skin. *New England Journal of Medicine*, **323**, 789–795.

Ha, I. D., Lee, Y., and MacKenzie, G. (2007) Model selection for multicomponent frailty models. *Statistics in Medicine*, **26**, 4790-4807.

Hodges, J. and Sargent, D. (2001) Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika*, **88**, 367–379.

Liang, H., Wu, H. L., and Zou, G. H. (2008) A note on conditional AIC for linear mixed-effects models. *Biometrika*, **95**, 773–778.

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

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, **92**, 351–370.

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

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, **93**, 120–131.

Growth curve modelling of a latent time-dependent risk factor in a multi-state model for stroke

Ardo van den Hout[1], Jean-Paul Fox[2], Rinke H. Klein Entink[2]

[1] MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 0SR, U.K. E-mail: ardo.vandenhout@mrc-bsu.cam.ac.uk.
[2] Department of Research Methodology, Measurement, and Data Analysis, Faculty of Behavioral Sciences, Twente University, The Netherlands.

Abstract: A three-state illness-death model is presented where transition intensities can be linked to time-dependent risk factors. State one is the healthy state, state two denotes a history of stroke, and state three is the death state. In addition to manifest factors, a time-dependent latent factor is added to the model. This latent factor is cognitive ability which is measured by a growth curve item response model for data obtained using a health-related questionnaire. Combining a multi-state model with the latent growth curve model defines a new model which extends current statistical inference regarding disease progression and cognitive ability. The method is illustrated using data from the Medical Research Council Cognitive Function and Ageing Study in the UK.

Keywords: cognition; item-response theory; Markov chain Monte Carlo.

1 Introduction

A continuous-time multi-state model can be used to describe disease progression over time. It is possible to link an intensity of moving from one state to another to risk factors for the disease such as age or sex.
We propose a model that links a latent time-dependent risk factor to transition intensities, where the change of the factor is described by a random-effects growth curve model. Specific to the application, the factor is latent cognitive ability and we assume that it can be measured by an item response theory (IRT) model for longitudinal scores on a questionnaire. An IRT model is a way of relating the probability of a series of discrete values to a latent continuous variable. In our case, the values are responses to binary questions, and the continuous variable is the ability that is assumed to explain how well individuals perform.
A linear mixed model was included in an IRT model by Fox and Glas (2001). We use a similar framework to define the latent growth curve model, although we use different restrictions to identify the model. Combining the

longitudinal IRT model with a multi-state model has not been described before, and seems promising in scope.
In the study that motivated this investigation, there is longitudinal information on progression of cardiovascular diseases and information on cognitive ability as measured by the Mini-Mental State Examination (MMSE, Folstein *et al.*, 1975). It is of interest whether cognitive ability can be identified as a predictor of cardiovascular diseases. In the application we will investigate this using data on survival and stroke from the Medical Research Council Cognitive Function and Ageing Study (CFAS, `www.cfas.ac.uk`).
The proposed methodology consists of two steps. First, Markov chain Monte Carlo methods (MCMC) in WinBUGS are used to sample values from the posterior density of the IRT growth curve model. Second, the MCMC output is used for multiple imputation of the latent variable for cognitive ability and the parameters of the multi-state model are estimated using maximum likelihood and Rubin's rules (Rubin, 1987).

2 The multi-state model

Let the interval-censored multi-state data be given by $\mathbf{x}_1, ..., \mathbf{x}_N$, where N is the sample size. The trajectory of individual i is given by $\mathbf{x}_i = (x_{i1}, ..., x_{in_i})$, where n_i is the number of observed states, and state $x_{ij} \in \{1, ..., S\}$. Times of observation - not necessarily equidistant - are given by $t_{i1}, ..., t_{in_i}$, where $t_{i1} = 0$, for all i, denotes the start of the study. For individual i we have risk factor values at the observation times, i.e, $\mathbf{w}_i = (\mathbf{w}_{i1}, ..., \mathbf{w}_{in_i})$. For now we assume that the values of the factors are observed (manifest).
We will assume a piecewise-constant multi-state model where individual trajectories through the states are conditionally independent. Transition intensity $q_{rs}(t_{ij})$ for going from state r to s is regressed on factors using $\log[q_{rs}(t_{ij})] = \mathbf{c}_{rs}^{\top}\mathbf{w}_{ij}$. For individual i, the likelihood contribution is

$$p(\mathbf{x}_i|\mathbf{c}, \mathbf{w}_i) = p(x_{in_i}|x_{i,n_i-1}, \mathbf{c}, \mathbf{w}_{i,n_i-1}) \times ... \times p(x_{i2}|x_{i1}, \mathbf{c}, \mathbf{w}_{i1})p(x_{i1}|\mathbf{c}, \mathbf{w}_i),$$

where $\mathbf{c} = (\mathbf{c}_{12}, \mathbf{c}_{13}, ...)$. We condition on the first state by restricting $p(x_{i1}|\mathbf{c}, \mathbf{w}_i) = 1$. The remaining probabilities at the right-hand side are individual transition probabilities.
As implied by the above, we assume that given the current state and the current values of the factors, the distribution of the next state does not depend on the states visited before the current state. In addition, we assume that factor values are constant between consecutive observation times. Within each individually observed time interval $(t_{ij}, t_{i,j+1}]$, this defines a time-homogeneous Markov process. Exact death times and right-censoring at the end of the follow-up can be taken into account. This model is fairly standard in the literature on multi-state models (Kalbfleisch and Lawless, 1985). By using age as a piecewise constant time-dependent risk factor, possible dependence of transition intensities on changing time can be taken into account (Van den Hout and Matthews, 2009).

3 Latent growth curve model

If a risk factor in the multi-state model is continuous, non-deterministic, and time-dependent, then we propose to describe this factor by a growth curve model with random effects.
Following the notation above, let the values of the factor be given by $\theta_{i1}, ..., \theta_{in_i}$ for individual i at times $t_{i1}, ..., t_{in_i}$. The growth curve model is given by

$$\theta_{ij} = \eta_{1i} + \eta_{2i} t_{ij} + e_{ij}. \tag{1}$$

Specific to the application in this paper, the factor is cognitive ability which is latent since it is not directly observed but measured by a test (a questionnaire). At every observation time, the test consists of K binary items (questions). For individual i, the data for the IRT model are given by $\mathbf{y}_i = (\mathbf{y}_{i1}, ..., \mathbf{y}_{in_i})$ and $\mathbf{y}_{ij} = (y_{ij1}, ..., y_{ijK})$. The probability of individual i answering item k correctly at time t_{ij} given item parameters $a_1, ..., a_K$ and $b_1, ..., b_K$ is

$$p(y_{ijk} = 1 | \theta_{ij}, a_k, b_k) = \Phi(a_k \theta_{ij} - b_k), \tag{2}$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution.
To identify the model, we use parameter restrictions. We define a standard cross-sectional IRT model for the baseline, and define the linear trend in the growth curve model conditional on the baseline parameters. For the baseline this means

$$\begin{aligned} \eta_{1i} &\sim N(0,1) \\ e_{i1} &= 0 \\ y_{i1k} | \mathbf{a}, \mathbf{b}, \theta_{i1} = \eta_{1i} &\sim Bernoulli\Big(\Phi(a_k \theta_{i1} - b_k)\Big). \end{aligned}$$

For the follow-up ($j > 1$) we assume

$$\begin{aligned} \eta_{2i} | \eta_{1i}, \rho, \sigma_\nu &\sim N\left(\nu + \rho \eta_{1i} \sigma_\nu, \sigma_\nu^2 (1 - \rho^2)\right) \\ e_{ij} | \sigma &\sim N(0, \sigma^2) \\ \theta_{ij} | \eta_{1i}, \eta_{2i}, t_{ij}, \sigma &\sim N(\eta_{1i} + \eta_{2i} t_{ij}, \sigma^2) \\ y_{ijk} | \mathbf{a}, \mathbf{b}, \theta_{ij} &\sim Bernoulli\Big(\Phi(a_k \theta_{ij} - b_k)\Big). \end{aligned}$$

where ρ is the correlation between the intercept η_{1i} and the slope η_{2i}. The conditional distribution of η_{2i} follows from the general formula for the conditional distribution of $Z_1 | Z_2 = z_2$ when both Z_1 and Z_2 are normally distributed.
With the restriction on baseline ability (θ_{i1}, $i = 1, ..., N$), both the baseline ability and the item parameter vectors $\mathbf{a}$ and $\mathbf{b}$ can be estimated using

baseline test scores y_{i1k}, $i = 1, ..., N$, $k = 1, ..., K$. This is a standard cross-sectional IRT model, where likelihood contributions for individual i are given by

$$p(\mathbf{y}_{i1}|\theta_{i1}, \mathbf{a}, \mathbf{b}) = \prod_{k=1}^{K} \Phi(a_k\theta_{i1} - b_k)^{y_{i1k}} \Big(1 - \Phi(a_k\theta_{i1} - b_k)\Big)^{1-y_{i1k}}.$$

Conditional on baseline ability and the item parameter vectors, the other parameters of the growth curve model can be estimated. In the implementation of this scheme, care has to be taken such that the follow-up does not inform parameter vectors **a** and **b** of the cross-sectional IRT model.
The resulting latent growth curve IRT model consists of two parts, namely the growth curve modeling (1) of the occasion-specific measurements, and the occasion-specific measurement (2) of the factor θ. This can be seen as a multi-group model where the baseline is the reference-group.
The latent growth curve IRT model is combined with the multi-state model via the regression equations for the transition intensities: $\log[q_{rs}(t_{ij})] = \mathbf{c}_{rs}^{\top}\mathbf{w}_{ij} + d_{rs}\theta_{ij}$.

4 Estimation

For the estimation of the IRT growth curve model, Bayesian inference is applied. Vague priors are used for the model parameters **a** **b**, ν, σ, σ_ν, and ρ. The model was implemented and estimated in WinBUGS using two MCMC chains. After convergence, sampled values for the latent ability θ were saved and used for multiple imputation. We used $M = 10$ data sets in which the imputed θ was combined with the corresponding three-state survival data. For each data set we fitted a three-state model using the R package msm (Jackson *et al.*, 2003) which performs maximum likelihood estimation. The results were then combined using Rubin's rules.

5 Summary of results

CFAS is a longitudinal population based study (1991-2004). All individuals at baseline are aged 65 years and above, and all deaths up to the end of 2005 have been included. In what follows we analyse the data for men in rural Cambridgeshire who have up to 4 observed living states, and - in addition - a death state or a censored state. $N = 1748$. Men are in state one when they do not have a history of stroke, and in state two when they do. State 3 is the death state. There are 499 men who only have one observed living state (at baseline).
In the three-state model without recovery time t is years since baseline. Transition intensities $q_{12}(t)$, $q_{13}(t)$, and $q_{23}(t)$ are regressed on an intercept,

t

TABLE 1. Parameter estimates (SEs) for the three-state model. Based on multiple imputed cognitive ability (θ).

Intercept			Cognitive ability (θ)		
$c_{12.0}$	-4.191	(0.147)	d_{12}	-0.606	(0.232)
$c_{13.0}$	-3.322	(0.068)	d_{12}	-0.310	(0.075)
$c_{23.0}$	-2.093	(0.134)	d_{12}	-0.192	(0.077)
Age					
$c_{12.1}$	0.028	(0.027)			
$c_{13.1}$	0.081	(0.009)			
$c_{23.1}$	0.046	(0.015)			

time-dependent (centred) age, and latent cognitive ability (θ_{ij} as defined above, with regression coefficients d_{12}, d_{13}, and d_{23}).

Table 1 presents the parameters of the three-state model for stroke history. As was to be expected, coefficients for age are positive showing that the risk of both a stroke and death increases with age. The sign of the coefficients for cognitive ability θ also concurs with expectations: decreasing cognitive ability is associate with an increase of the risk of a transition to a next state be it the ill health state or the death state.

As an example of posterior inference that might be of interest, we consider transition probabilities given a specific change of cognitive ability over time. For subject i, we assume that $\eta_{1i} = 0$ and that there is no change of ability over time (slope $\eta_{2i} = 0$). For subject h, we also assume that $\eta_{1h} = 0$, but the ability changes over time with slope $\eta_{2h} = \widehat{\nu} - \widehat{\sigma}_{\nu}$, where the estimates are the posterior means of the parameters. In words, subject h experiences a decline in cognitive ability parametrised by the location of the random slope minus one standard deviation. Given times $t_{i1}, ..., t_{in_i}$, for both i and h, mean cognitive ability over time can be inferred from the growth curve model.

Say both subjects are 65 years old at baseline $t_{i1} = 0$, and $t_{in_i} - t_{i1}$ is 10 years. Transition probabilities can be derived from the point estimates of the parameters of the three-state model if changing (and centred) age is taken into account piecewise-constantly. We get estimated probabilities $p(x_{in_i} = 2|x_{i1} = 1) = 0.068$ and $p(x_{hn_i} = 2|x_{h1} = 1) = 0.104$. This concise comparison shows that a decrease of cognitive ability over time is associated with an increase of the probability of a stroke.

Item parameters and the growth curve model parameters have not been discussed in the above as the focus of this report is on the multi-state model.

However, the longitudinal IRT model can be of interest as a separate model with regard to possible differences in cognitive decline between individuals. The Bayesian estimation using WinBUGS is not very fast, but it provides a flexible framework that can easily be extended to more complex models.

References

Folstein, M.F., Folstein, S.E., and McHugh, P.R. (1975). Mini-mental state. A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research* **12**, 189−198.

Fox, J.-P., and Glas, C.A.W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika* **66**, 271−288.

Jackson, C.H., Sharples, L.D., Thompson, S.G., Duffy S.W., and Couto E. (2003). Multi-state Markov models for disease progression with classification error. *Statistician* **52**, 193−209.

Kalbfleisch, J. and Lawless, J.F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association* **80**, 863−871.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Van den Hout A., Matthews F.E. (2009). Estimating dementia-free life expectancy for Parkinson's patients using Bayesian inference and microsimulation. *Biostatistics* **10**, 729−743.

Spatiotemporal smoothing of brain magnetoencephalography data

M. Ventrucci[1], A. W. Bowman[1], C. Ferguson[1], J. Gross[2], J. M. Schoffelen[3]

[1] Department of Statistics, University of Glasgow, Glasgow G12 8QQ, UK
[2] Department of Psychology, University of Glasgow, Glasgow G12 8QQ, UK
[3] Donders Institute for Brain, Cognition and Behaviour, Centre for Cognitive Neuroimaging, Radboud University Nijmegen, Nijmegen, The Netherlands

Abstract: Magnetoencephalography (MEG) is a non-invasive technique to measure the neuronal activity in the brain. An electromagnetic field is measured at a very high temporal resolution in many sensors located in a helmet-shaped dewar producing a very large dataset. In this paper spatiotemporal smoothing is applied in order to identify spatial and temporal patterns and to build a gradient estimator helpful in suggesting the locations of the current dipoles in the brain. Results when applying these tools to a MEG experiment are shown.

Keywords: spatiotemporal smoothing; MEG; current dipole.

1 Introduction

Magnetoencephalography (MEG) is a non-invasive technique which measures the electromagnetic activity in the brain by recording the magnetic fields outside the head. Data are acquired by sensitive devices (magnetometers) embedded in a helmet placed over the human skull. In general one or more subjects are involved in multiple replicates of an experiment where, by means of a stimulus, the subject is asked to do a particular task and the MEG signal is measured before, during and after the stimulus. The MEG data are available at a very high temporal resolution (order of msec) for many sensors placed across the helmet. Several kinds of noise and artifacts can distort the desired signal. Filtering procedures and averaging across replicates are methods usually adopted to increase the signal to noise ratio.

In this paper the benefit of applying linear smoothing estimators to the MEG data is investigated. Smoothed estimates can depict the mean spatiotemporal surface of the signal and standard errors can be computed for inferential purposes. A planar gradient estimate of the smoothed signal is obtained as a means of identifying the location of the neuronal sources in the brain where activity takes place. Neuronal source activity is usually modelled by one or more current dipoles which represent the primary cur-

rents associated with the activation of a large number of neurons located in a small region. The problem of identifying current dipoles from the MEG signal measured at the scalp, often referred as the inverse problem, has no unique deterministic solution and has been addressed in the statistical literature; see the seminal paper by Mosher et al. (1992)
A dipole effect in the MEG signal is identified by two close regions in the map showing very high but opposite sign fields. Thus a gradient map depicting the MEG signal gradients across the surface might help in approximating the location of possible current dipoles, and a gradient peak should occur where a current dipole is operating. A gradient computation can be achieved by differentiating the signal obtained from spatiotemporal smoothing over an estimation grid superimposed over the brain surface.

1.1 The motivating example

The data are collected from replicates of an MEG experiment conducted on 19 subjects to study the event-related response in 5 different stimulus configurations. For one replicate the MEG signal was recorded on $S = 248$ sensors and $T = 256$ time points and spans a time window from half a second before to half a second after the stimulus. Here interest is focused on only one subject in one stimulus configuration, that is a red light appearing from the left indicating to press the left hand button, for which 135 replicates are available. A first activation is expected in response to the light stimulus (to acknowledge a light from the left) and a second one in response to the task requested (to press the left hand button). Both the replicate level signal and the signal averaged across replicates were analysed.

2 Methods

The model assumed is:

$$E[y_{st}] = m(s,t), \quad \forall\ s = 1, ..., S; t = 1, ..., T,$$

where m is a smooth function, defining a local mean spatiotemporal surface, $m(s,t)$, for each measurement y_{st} taken in sensor s and time point t. A first issue is the difficulty of applying standard smoothing techniques across the 3d helmet surface. In this spatial domain an appropriate weight function for the sensor observations at each time point is based on the geodesic distances between each sensor $\{d_{jk}; j, k = 1, ..., S\}$. Once geodesic distances are calculated a normal kernel density $\exp(-0.5d_{jk}^2/h^2)$ can be used to construct the weights for local averaging.
A computationally efficient solution for $\hat{m}(s,t)$ can be obtained by exploiting the array representation of the data $Y_{S\times T}$ and smoothing the data by marginal operations by means of S_s and S_t, the space and time smoothing matrices respectively; see Currie et al. (2006) for further details of

this general approach, and Bowman et al. (2009) for an application to spatiotemporal smoothing of environmental data. A local mean estimator can be calculated for each point of a spatiotemporal arbitrary grid as $\widehat{m} = S_s Y {S_t}^T$. Standard errors are obtained by assuming a separable model for the variance-covariance of the process, $\Sigma_y = (R_t \otimes R_s)\,\sigma^2$. An estimate of R_t is obtained by fitting an AR(1) model to temporal residuals for each sensor location, while R_s is estimated by fitting an exponential model to the empirical variogram worked out for each time point, and σ^2 can be estimated as an average of the error variances obtained from the temporal and spatial residual analysis.

As discussed above a gradient representation of the MEG signal can inform us about the rate of change of the MEG field across the helmet. To calculate the gradients at each point of the spatiotemporal grid, derivatives of the smooth estimated signal were computed in both horizontal (β_x) and vertical (β_y) directions and transformed as $\Delta_{st} = \sqrt{{\beta_x}^2 + {\beta_y}^2}$, which identifies the highest rate of change. Again standard errors can be provided in order to investigate the significance of slopes across the brain.

3 Summary

In Figure 1 results from the spatiotemporal smoothing at replicates 56 and 135, and averaged across replicates, are shown at 360 msecs after the stimulus. The 3d brain surfaces are flattened here in a 2d representation. After 360 msecs a clear response to the stimulus should take place and brain activation related to the task of pressing the left hand button should be detected. The left hand panels present the raw signal and highlight spatial variation but unclear dipole effects, especially at the replicate level. In the central panel the results after smoothing are displayed and the dipole effects are clearer than before, showing in a more pronounced manner the paired red and yellow regions with opposite sign fields. For example at the average level a dipole effect seems to appear in the left bottom of the map, suggesting a brain region where a source of activation may be located. Looking at the gradient map in the panels on the right, a current dipole is more accurately identified through the small blue region.

3.1 Discussion

The methodology discussed here offers potential for the detection of spatiotemporal patterns of brain activation. It also allows inferential methods, for instance, to compare the spatiotemporal surface in different conditions, such as between left hand and right hand button pressing tasks. In addition inferences on gradients offer an attractive way to approximate the source locations, for instance by means of multiple testing of a null gradient across the map. Such a gradient approach can provide a preliminary solution to the source localization problem and be a basis of further analysis.

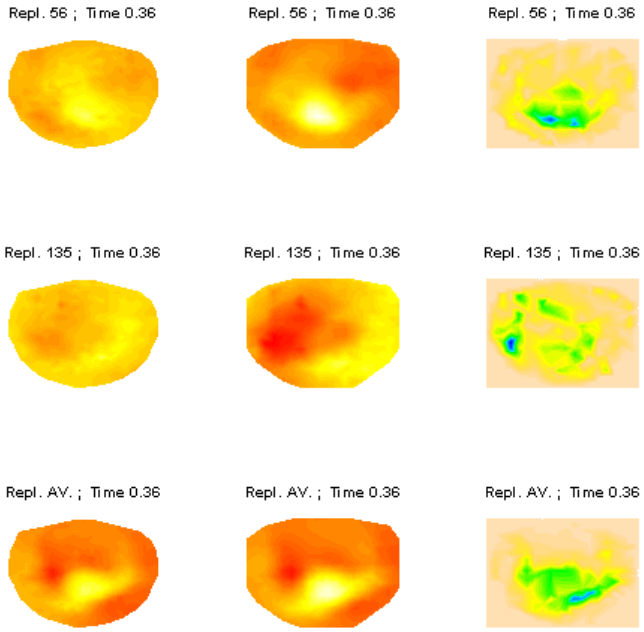


FIGURE 1. MEG spatiotemporal smoothing and gradient estimation for singular replicates and averaged across replicates signal (AV.) at a given snapshot (360 msecs after the stimulus). On the left hand panels the MEG signal (red indicates positive field, yellow indicates negative field), in the central panels the smoothed signal and on the right hand panels the gradients (in blue the highest slopes suggesting possible current dipoles).

References

Bowman, A.W., Giannitrapani, M., and Scott, E.M. (2009). Spatiotemporal modelling and sulphur dioxide trends over Europe. *Journal of the Royal Statistical Society, Series C, Applied Statistics*, **58**, 737-752.

Currie, I.D., Durban, M., and Eilers, P.H.C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society, Series B*, **68**, 259-280.

Mosher, J.C., Lewis, P.S., and Leahy, R.M. (1992). Multiple dipole modeling and localization from spatio-temporal MEG data. *IEEE Transaction on Biomedical Engineering*, **39**, 541-557.

Misspecification Effects in the Analysis of Longitudinal Survey Data

Marcel de Toledo Vieira[1], Maria de Fátima Salgueiro[2], Peter W.F. Smith[3]

[1] Universidade Federal de Juíz de Fora, Brazil, marcel.vieira@ufjf.edu.br
[2] ISCTE-IUL - Lisbon University Institute, Portugal, fatima.salgueiro@iscte.pt
[3] S^{3}RI, University of Southampton, UK, p.w.smith@soton.ac.uk

Abstract: Misspecification effects (*meff*s) measure the inflation of the sampling variance of an estimator as a result of the use of complex sampling schemes. Many longitudinal social survey designs employ multi-stage sampling, leading to some clustering of the sample and to *meff*s greater than one. For a model for panel data we consider methods for estimating parameters which allow for complex schemes. An empirical study using longitudinal data from the British Household Panel Survey is conducted, and a simulation study is performed.

Keywords: Parametric models; Longitudinal data; Sampling impacts.

1 Introduction

Standard inferential methods are often not valid when analysing data obtained using a complex sampling scheme. The interest in fitting models to longitudinal complex survey data has been growing in the last decade. Skinner and Vieira (2007) presented evidence that the variance-inflating impacts of clustering may be higher for longitudinal analyses than for the corresponding cross-sectional analyses. We further investigate the impact of weighting, stratification and clustering in the regression analysis of longitudinal survey data, comparing it with the impact on cross-sectional analyses. In Section 2 we introduce the longitudinal survey data under analysis. Section 3 presents the model, point and variance estimation procedures, and describes measures of misspecification effects (*meff*s). The motivating application and empirical results are presented in Section 4 and a simulation study is performed in Section 5. Section 6 contains a discussion.

2 Data and Sampling Design

The empirical evidence presented in this paper is based on data from the British Household Panel Survey (BHPS), a household panel survey of individuals in private domiciles in Great Britain. The BHPS follows longitudinally a sample of individuals selected in 1991 by a complex stratified

two-stage sampling scheme, with clustering by area. Our analyses are based on a subsample of 2255 men and women aged 16 or more, who were original sample members, who gave a full interview in waves twelve to fifteen, and who were employed throughout the period. The following variables are considered: gender; age category; number of children in the household; qualification; social class; marital status; health status; hours normally worked per week; and logarithm of the household income. In our sample, the relative frequency for both gender categories is approximately 50%. The distribution of the age category variable is negatively skewed, as the frequencies for the older categories are larger. Most of the respondents are either married or living as a couple in 2002. Approximately 80% of the respondents considered themselves in either good or excellent health condition. Furthermore, over 75% of the individuals worked at least 30 hours per week. About 55% of the individuals had a high level of education, and only 16.32% of them occupied a partly skilled or an unskilled position in their last job. Almost 62% of the respondents had no children in the household where they live. Moreover, the average household income of the sample members was approximately £3365 in the month before the interview was made.

3 Model, Estimation Procedures and *Meff*s

Regression models have found a wide range of useful applications with longitudinal survey data (e.g., Diggle et al. 2002; Vieira, 2009). Let y_{it} denote the response of interest for individual i at time t. Let $y_i = (y_1, \ldots, y_{iT})'$ be the vector of repeated measures. We consider linear models of the following form to represent the expectation of y_i given the values of covariates:

$$E(y_i) = x_i\beta, \tag{1}$$

where $x_i = (x'_{i1}, \ldots, x'_{iT})'$, x_{it} is a $1 \times q$ vector of covariates for individual i at wave t, β is the $q \times 1$ vector of regression coefficients.
Following the pseudo-likelihood approach (Skinner, 1989), the most general estimator of β considered in this paper is

$$\hat{\beta} = \left(\sum_{i\in s} w_i x'_i V^{-1} x_i\right)^{-1} \sum_{i\in s} w_i x'_i V^{-1} y_i = A^{-1} \sum_{i\in s} w_i x'_i V^{-1} y_i, \tag{2}$$

where w_i is a longitudinal survey weight, V is a $T \times T$ estimated 'working' variance matrix of y_i (Diggle *et al.* 2002), taken as the exchangeable variance matrix with diagonal elements $\hat{\sigma}^2$ and off-diagonal elements $\hat{\rho}\hat{\sigma}^2$. Under (1), $\hat{\beta}$ is approximately unbiased with respect to the model and the survey design, and is expected to combine both within and between individual information in a reasonably efficient manner, even if the working model for the error structure does not hold exactly. Without the weight terms

and survey sampling considerations, the form of $\hat{\beta}$, given by (2), is motivated by the generalized estimating equations (GEE) approach of Liang and Zeger (1986), which we shall denote by $\hat{\beta}_n$. The following estimator of the covariance matrix of $\hat{\beta}$, based on linearization (Skinner, 1989; Skinner and Vieira, 2007), allows for a stratified multistage sampling scheme:

$$v(\hat{\beta}) = A^{-1}\left[\sum_h n_h/(n_h-1)\sum_a (z_{ha}-\bar{z}_h)(z_{ha}-\bar{z}_h)'\right]A^{-1},$$

where h denotes stratum, a denotes primary sampling unit (PSU), n_h is the number of PSUs in stratum h, $z_{ha} = \sum_i w_i x_i' V^{-1} e_i$, $\bar{z}_h = \sum_a z_{ha}/n_h$ and $e_i = y_i - x_i\hat{\beta}$. We consider three further alternatives for estimating the covariance matrix of $\hat{\beta}$: (i) $v_a(\hat{\beta})$, which considers that the population consists of only one stratum (h =1), and therefore ignores stratification; (ii) $v_h(\hat{\beta})$, which considers that each individual i is a PSU, and therefore ignores clustering; and (iii) $v_n(\hat{\beta})$, which considers that h =1 and that each individual is a PSU, and therefore ignores both stratification and clustering. We also perform variance estimation for $\hat{\beta}_n$. We are concerned with the potential bias of $v_a(\hat{\beta})$, $v_h(\hat{\beta})$, and $v_n(\hat{\beta})$, when in fact the design is complex. Skinner (1989) has proposed the *meff*, which is designed to measure the effects of incorrect specification of both the sampling scheme and the considered model. We consider $meff_a[\hat{\beta}_k, v_a(\hat{\beta}_k)] = v(\hat{\beta}_k)/v_a(\hat{\beta}_k)$; $meff_h[\hat{\beta}_k, v_h(\hat{\beta}_k)] = v(\hat{\beta}_k)/v_h(\hat{\beta}_k)$; and $meff_n[\hat{\beta}_k, v_n(\hat{\beta}_k)] = v(\hat{\beta}_k)/v_n(\hat{\beta}_k)$, where $\hat{\beta}_k$ denote the k^{th} element of $\hat{\beta}$. The $meff_a$, $meff_h$, and $meff_n$ measure the impact of stratification, clustering, and both stratification and clustering, respectively. We also estimate all the considered versions of the *meff* for $\hat{\beta}_n$. Furthermore, $meff_g = v(\hat{\beta}_k)/v_n(\hat{\beta}_{nk})$ is estimated in order to access the bias caused by ignoring all the features of the sampling scheme.

4 Application

The paper is motivated by a regression analysis of four waves of BHPS data, which considers logarithm of the household income as our dependent variable. We first estimate *meff*s for the linearization estimator, considering $\hat{\beta}$, as discussed in Section 3. Using data from just the first wave and setting x_i =1, the estimated $meff_n$ for the cross-sectional mean is given in Table 1 as about 1.3. In order to evaluate the impact of the longitudinal aspect of the data, we estimated a series of each type of the *meff*s discussed above, using data for waves 12 to 15. Although these estimated *meff*s are subject to sampling error, there is a tendency for $meff_h$, $meff_n$, and $meff_g$ to increase with the number of waves (Table 1). It therefore seems that it becomes more important to allow for clustering and for the complex sampling design in general when the number of waves in the analysis increases. Furthermore,

TABLE 1. *Meff* estimates for longitudinal means.

Meff	Waves			
	12	12 and 13	12 to 14	12 to 15
$meff_a[\hat{\beta}_k, v_a(\hat{\beta}_k)]$	0.971	0.965	0.965	0.963
$meff_h[\hat{\beta}_k, v_h(\hat{\beta}_k)]$	1.490	1.653	1.699	1.695
$meff_n[\hat{\beta}_k, v_n(\hat{\beta}_k)]$	1.282	1.431	1.474	1.458
$meff_a[\hat{\beta}_{nk}, v_a(\hat{\beta}_{nk})]$	0.969	0.963	0.961	0.960
$meff_h[\hat{\beta}_{nk}, v_h(\hat{\beta}_{nk})]$	1.572	1.795	1.830	1.870
$meff_n[\hat{\beta}_{nk}, v_n(\hat{\beta}_{nk})]$	1.343	1.504	1.575	1.653
$meff_g$	1.494	1.598	1.778	1.706

stratification effects appear to be constant with increases in the number of waves.

When we included educational level as a covariate, we also noticed some evidence for $meff_h$, $meff_n$, and $meff_g$ to increase with the number of waves. The model has been further elaborated by adding time, gender, age category, marital status, number of children in the household, social class, health status and numbers of hours normally worked per week as covariates. Once more, we observed some evidence of a tendency for those *meff*s to diverge from one as the number of waves increases, at least for the coefficients of some of the covariates. We also confirmed the observation of Skinner and Vieira (2007) that *meff*s for regression coefficients tend not to be greater than *meff*s for the means of the response.

5 Simulation Study

As results reported in Section 4 are subject to sampling error we have conducted a simulation study to evaluate the behaviour of the *meff* measures. Each of the $d = 1, \ldots, D$ replicate samples is based on the BHPS data subset described above which is considered as the 'target population'. We evaluated the properties of variance estimators for unweighted point estimators and assessed only different impacts of clustering. We studied the *meff* when the number of waves in the analysis is increased. Note that we did not assess the impact of either stratification or unequal probability sampling.

Let y_{iat} be the value for the study variable for unit $i = 1, 2, \ldots, n_d^{sim}$, in PSU $a = 1, \ldots, m_d^{sim}$, at wave t of the survey, where n_d^{sim} and m_d^{sim} are the sample size and the number of PSUs for the replicate sample d. For generating the values of y_{iat} for the simulation study, we used the following uniform correlation model which allows for the impact of clustering:

$$y_{iat} = x_{iat}\beta + \eta_a + u_{ia} + v_{iat}, \tag{3}$$

with $\eta_a \sim N(0, \sigma_\eta^2)$, $u_{ia} \sim N(0, \sigma_u^2)$, and $v_{iat} \sim N(0, \sigma_v^2)$. We have considered the logarithm of the household income as the dependent variable and the remaining variables listed in Section 2 as covariates. We have held the values of the covariates fixed.

The adopted values for β, σ_η^2, σ_u^2, and σ_v^2 have been obtained by maximum likelihood estimation considering the 'target population'. In particular, we have considered different realistic choices for σ_η^2, namely $\sigma_\eta^2 = 0.06$ (actual value estimated from fitting the model given by (3)), $\sigma_\eta^2 = 0.12$, and $\sigma_\eta^2 = 0.18$ to enable the evaluation of effects of different amounts of clustering on the considered variance estimation procedures.

Let

$$\hat{E}(\hat{meff}) = \frac{1}{D}\sum_{d=1}^{D} \hat{meff}^{(d)},$$

be the mean of our parameter of interest estimated over repeated simulation,

$$var(\hat{meff}) = \frac{1}{D-1}\sum_{d=1}^{D}\left[\hat{meff}^{(d)} - \hat{E}(\hat{meff})\right]^2,$$

be a simulation estimator of $VAR(\hat{meff})$, the population variance of the misspecification effect measure, and

$$se\left[\hat{E}(\hat{meff})\right] = \sqrt{var(\hat{meff})/D}$$

be the simulation standard error of $\hat{E}(\hat{meff})$.

For the models that have been fitted to each generated replicate sample, we have set x_i =1 and therefore we have still only studied the behavior of the *meff* for longitudinal means. Let n_a be the sample size for PSU a in the 'target population' and n_{da}^{sim} be the sample size for PSU a in the replicate sample d.

Table 2 presents the results for three scenarios: (i) ($m_d^{sim} = 200$, $n_{da}^{sim} = n_a$, and $\sigma_a^2 = 0.06$); (ii) ($m_d^{sim} = 200$, $n_{da}^{sim} = n_a$, and $\sigma_a^2 = 0.12$), and (iii) ($m_d^{sim} = 200$, $n_{da}^{sim} = n_a$, and $\sigma_a^2 = 0.18$). Note that $m = 234$ in the 'target population'.

The simulation results also give evidence that there is a tendency for the *meff* to increase as the number of waves in the analysis increases, at least for longitudinal means. This tendency seems to be stronger for larger clustering impacts. *Meff*s also increase when the amount of clustering is increased, as expected from the survey sampling literature (Vieira, 2009). Simulation standard errors for $\hat{E}(\hat{meff})$ appear to increase when the number of waves and the amount of clustering are increased.

TABLE 2. $\hat{E}(\hat{meff})$ and $se\left[\hat{E}(\hat{meff})\right]$ (in brackets), for three scenarios.

$n_j^{sim^*}$	σ_η^2	Waves			
		12	12 and 13	12 to 14	12 to 15
n_j	0.06	1.1901	1.2077	1.2115	1.2143
		(0.0044)	(0.0046)	(0.0047)	(0.0047)
	0.12	1.2766	1.3014	1.3106	1.3157
		(0.0054)	(0.0057)	(0.0058)	(0.0058)
	0.18	1.3624	1.3933	1.4061	1.4118
		(0.0066)	(0.0069)	(0.0070)	(0.0070)

D=1000

6 Discussion

We have presented some evidence that clustering impacts may be stronger for longitudinal studies than for cross-sectional studies, and that *meff*s for the regression coefficients may increase with the number of waves considered in the analysis. The main implication of these findings is that standard errors in the analysis of longitudinal survey data may be misleading if the initial sample was clustered and if this clustering is ignored. We have also observed that *meff*s for regression coefficients tend not to be greater than *meff*s for the means of the dependent variable.

Acknowledgments: The research of the first author was supported by the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) grant CEX-APQ-00467-2008. The research of the second author was supported by the Portuguese FCT grant PTDC/GES/72784/2006.

References

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

Liang, K. and Zeger, S. L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, **73**, 13-22.

Skinner, C.J. (1986). Domain means, regression and multivariate analysis. In: *Analysis of Complex Surveys*, Skinner, C. J., Holt, D. and Smith, T. M. F. (eds.), 59-87. Chichester: Wiley.

Skinner, C. and Vieira, M. D. T. (2007). Variance estimation in the analysis of clustered longitudinal survey data. *Survey Methodology*, **33**, 3-12.

Vieira, M. D. T. (2009). *Analysis of Longitudinal Survey Data*. Saarbrücken: VDM Verlag Dr. Müller.

Bandwidth Matrix Choice for Bivariate Kernel Density Derivative

Kamila Vopatová[1,2], Ivanka Horová[1], Jan Koláček[1]

[1] Dept. of Mathematics and Statistics, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic, vopatova@mail.muni.cz
[2] Dept. of Statistics and Operation Analysis, Mendel University in Brno, Czech Republic

Abstract: The aim of this contribution is to develop a method for bandwidth matrix choice for bivariate kernel density derivative.

Keywords: Kernel; bandwidth matrix; gradient estimator; convolution.

1 Introduction

Kernel estimate belongs to very effective nonparametric estimates of both probability density and its derivatives. There is a vast literature on univariate kernel density estimates, see e.g. Scott (1992), Wand & Jones (1995), Bowman & Azzalini (2003). But the progress in a multivariate case is much slower. The basic problem, which determines the performance of multivariate kernel density estimate, is the bandwidth matrix selection. General bandwidth matrix choices are described in Duong & Hazelton (2005), Chacón *et al.* (2009). Significant information about features of the density is contained in the first and the second derivative of the true density f. From this reason we focus on kernel density derivative estimator of the first derivative ∇f of f and the bandwidth matrix selection in bivariate kernel estimate of ∇f provided that the bandwidth matrix is diagonal.

2 Density estimate

First, we give a short overview of the bivariate kernel density estimate. Consider a bivariate random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ coming from an unknown density f. For this random sample the kernel density estimator is defined

$$\hat{f}(\mathbf{x}, H) = \frac{1}{n} \sum_{i=1}^{n} K_H\big(\mathbf{x} - \mathbf{X}_i\big)$$

where H is a symmetric positive definite 2×2 matrix called the bandwidth matrix, $K_H(\mathbf{z}) = |H|^{-1/2} K(H^{-1/2}\mathbf{z})$, where $|H|$ stands for the determinant of H, and K is a bivariate kernel function.

The kernel function K is often taken to be a probability density function. The consequence of this assumption is that the estimate $\hat{f}(\mathbf{x}, H)$ is a density function. There are two common ways for generating the two-dimensional kernel K from the one-dimensional kernel k: product kernel $K^P(\mathbf{x}) = k(x_1) \cdot k(x_2)$ and spherically symmetric kernel $K^S(\mathbf{x}) = c^{-1}k(\sqrt{x_1^2 + x_2^2})$, $c = \int\int k(\sqrt{x_1^2 + x_2^2})\,\mathrm{d}x_1\mathrm{d}x_2$. It is well known that the choice of K is not crucial, but much more important is the choice of the bandwidth matrix H, which controls the smoothness of the estimate $\hat{f}$. In this contribution we assume that H is a diagonal matrix $H = diag(h_1^2, h_2^2)$. A class of such matrices is denoted by $\mathcal{H}_{\mathcal{D}}$. The quality of the estimate $\hat{f}(\cdot\,, H)$ can be expressed in terms of mean integrated square error (MISE)

$$MISE(H) \equiv MISE\big(\hat{f}(\cdot\,, H)\big) = \int E\big(\hat{f}(\mathbf{x}, H) - f(\mathbf{x})\big)^2\,\mathrm{d}\mathbf{x}.$$

Because MISE is not mathematically tractable, we can go over to an asymptotic mean integrated square error (AMISE) under few assumptions about the kernel K, the bandwidth matrix H and the density f (see e.g. Wand & Jones (1995)). Horová *et al.* (2008) developed an iterative method for the bandwidth matrix choice.

3 Density derivative estimate

The kernel density derivative estimator of the first derivative ∇f of f is

$$\widehat{\nabla f}(\mathbf{x}, H) = \frac{1}{n}\sum_{i=1}^{n} \nabla K_H(\mathbf{x} - \mathbf{X}_i),$$

see Duong *et al.* (2008), where ∇f is column vector of the first partial derivatives $\nabla f = (\partial_1 f, \partial_2 f)^T$ and $\nabla K_H(\mathbf{z}) = |H|^{-1/2}H^{-1/2}\nabla K(H^{-1/2}\mathbf{z})$. For a gradient estimator $\widehat{\nabla f}(\mathbf{x}, H)$ the MISE is a matrix. Since a performance based on scalar quantities rather than on matrices is easier, it is appropriate to apply a matrix norm. We use the trace norm (Duong *et al.* (2008)) and the trace of asymptotic mean integrated square error (TAMISE$^{(1)}$) can be expressed as a sum of the trace of integrated variance (TIVar) and the trace of integrated square bias (TIBias^2). For the sake of simplicity we denote

$$D = \psi_{6,0} + \psi_{4,2}, \qquad E = \psi_{4,2} + \psi_{2,4}, \qquad F = \psi_{2,4} + \psi_{0,6},$$

where $\psi_{k,\ell} = \int \Big(\frac{\partial^3 f}{\partial x_1^{k/2}\partial x_2^{\ell/2}}\Big)^2\,\mathrm{d}\mathbf{x}$, $k, \ell = 0, 2, 4, 6$, $k+\ell = 6$. Then TAMISE$^{(1)}$ is given by

$$TAMISE^{(1)}(H) \equiv TAMISE^{(1)}\big(\widehat{\nabla f}(\cdot\,, H)\big) = TIVar + TIBias^2 =$$

$$= \frac{1}{nh_1h_2}\Big(\frac{V(\partial_1 K)}{h_1^2} + \frac{V(\partial_2 K)}{h_2^2}\Big) + \frac{1}{4}\beta_2(K)^2\big(h_1^4 D + 2h_1^2h_2^2E + h_2^4F\big)$$

$V(\partial_i K) = \int\Big(\frac{\partial K(\mathbf{x})}{\partial x_i}\Big)^2 \mathrm{d}\mathbf{x}$ and $\beta_2(K) = \int x_i^2 K(x_i)\,\mathrm{d}x_i$ is independent on i $(i = 1, 2)$. Under some additional assumptions on K and f the asymptotic properties of $\widehat{\nabla f}$ has been studied in Duong *et al.* (2008).
Let us consider the product Epanechnikov kernel and thus $V(\partial_1 K) = V(\partial_2 K)$ and we denote this quantity as Q. In such a case TAMISE$^{(1)}$ takes the form

$$TAMISE^{(1)}(H) = \frac{Q}{nh_1h_2}\Big(\frac{1}{h_1^2} + \frac{1}{h_2^2}\Big) + \frac{1}{4}\beta_2(K)^2(h_1^4D + 2h_1^2h_2^2E + h_2^4F).$$

Now, we make a specific assumption $h_2 = ch_1$ (see e.g. Scott (1992), Bowman & Azzalini (2003)). It significantly simplifies $TAMISE^{(1)}$

$$TAMISE^{(1)}(H) = \frac{Q(1+c^2)}{nc^3h_1^4} + \frac{h_1^4}{4}\beta_2(K)^2(D + 2c^2E + c^4F).$$

and offers a possibility to detect an optimal value of h_1

$$h_{1,opt}^8 = \frac{4Q(1+c^2)}{n\beta_2(K)^2c^3(D + 2c^2E + c^4F)}$$

i.e. $h_{1,opt}^2 \approx \mathcal{O}(n^{-1/4})$, which corresponds the result of Duong *et al.* (2008), Chacón *et al.* (2009). Moreover $TAMISE^{(1)}(H_{opt}) = \mathcal{O}(n^{-1/2})$, where $H_{opt} = diag(h_{1,opt}^2, c^2h_{1,opt}^2)$.
Further, for the TIVar and TIBias^2

$$TIVar\big(\widehat{\nabla f}(H_{opt})\big) = TIBias^2\big(\widehat{\nabla f}(H_{opt})\big) \qquad (\circ)$$

is valid. The relation $(\circ)$ serves as a basis of a method we are going to present. Our aim is to find h_1 such that $(\circ)$ is satisfied. But TIVar and TIBias^2 depend on the unknown density f through $\psi_{k,\ell}$. In order to avoid this dependence we use suitable estimates of TIVar and TIBias^2. It can be shown that ($*$ denotes a convolution)

$$\begin{aligned}
\widehat{TIVar}\big(\widehat{\nabla f}(H)\big) &= \mathrm{tr}\Big\{\tfrac{1}{n}|H|^{-1/2}H^{-1/2}V(\nabla K)H^{-1/2}\Big\},\\
\widehat{TIBias}^2\big(\widehat{\nabla f}(H)\big) &= \mathrm{tr}\Big\{\tfrac{1}{n^2}\sum_{i,j=1}^{n}\int\big[(K_H * \nabla K_H - \nabla K_H)(\mathbf{x} - \mathbf{X}_i)\big]\\
&\qquad \times \big[(K_H * \nabla K_H - \nabla K_H)(\mathbf{x} - \mathbf{X}_j)\big]^T\,\mathrm{d}\mathbf{x}\Big\}.
\end{aligned}$$

Taking these estimates into account and using $(\circ)$, we arrive at the nonlinear equation for h_1, $(h_2 = ch_1)$

$$n(1+c^2)Q + c^2A + B = 0,$$

where $A = A(h_1, K, \mathbf{X}_i)$ and $B = B(h_1, K, \mathbf{X}_i)$. Figure 1 explains the idea of developed method.

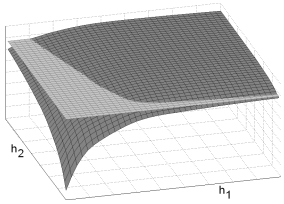


FIGURE 1. $\widehat{TIBias}^2\big(\widehat{\nabla f}(H)\big) - \widehat{TIVar}\big(\widehat{\nabla f}(H)\big) = 0.$

4 Conclusion

The proposed method is easy to implement especially for product kernels and using the fast computations of convolutions.

Acknowledgments: This research was supported by the Ministry of Education, Youth and Sports of the Czech Republic under the project MŠMT LC06024.

References

Bowman, A.W., and Azzalini, A. (2003). Computational aspects of nonparametric smoothing with illustrations from the `sm` library. *Computational Statistics and Data Analysis*, **42**, 545-560.

Chacón, J.E., Duong, T., and Wand, M.P. (2009). Asymptotics for general multivariate kernel density derivative estimators. http://www.uow.edu.au/~mwand/papers.html.

Duong, T., and Hazelton, M.L. (2005). Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, **32**, 485-506.

Duong, T., Cowling, A., Koch, I., and Wand, M.P. (2008). Feature significance for multivariate kernel density estimation. *Computational Statistics and Data Analysis*, **52**, 4225-4242.

Horová, I., Koláček, J., Zelinka, J., and Vopatová, K. (2008). Bandwidth choice for kernel density estimates. In: *Proceedings of IASC 2008*, 542-551, Yokohama, Japan.

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: John Wiley & Sons.

Wand, M.P., and Jones, M.C. (1995). *Kernel Smoothing*, London: Chapman & Hall.

Bayesian variable selection with spike and slab priors in logit models

Helga Wagner[1], Christine Duller[1]

[1] Johannes-Kepler-University Linz, Altenbergerstrasse 69, 4040 Linz, Austria, e-mail: helga.wagner@jku.at (corresponding author)

Abstract: Two different types of spike and slab priors for Bayesian variable selection in logit type models are compared. For both priors variable selection is accomplished by MCMC sampling based on a new auxiliary mixture sampler for logit models. Performance of both versions is compared for simulated data with particular focus on the sampling efficiency of posterior inclusion probabilities. Both methods are applied to a data set where the goal is to identify risk factors for complications after ERCP (endoscopic retrograde cholangiopancreatography).

Keywords: spike and slab prior, MCMC, auxiliary mixture sampling

1 Introduction

In many medical studies the goal is to identify factors effecting risk of diseases or adverse treatment effects. Risk of an adverse event is usually modelled using a logit type model, with a large set of potential risk factors available as covariates. Bayesian variable selection methods are widely used to identify those regressors which have non-zero effects and should be included in the final model. Many of these methods use spike and slab priors for the regression coefficients with the spike around zero to allow shrinkage of small effects to zero and a flat slab elsewhere. Basically there are two types of spike, a Dirac spike at zero or a continuous spike. For these prior types Bayesian variable selection is implemented for logit models based on a new auxiliary mixture sampler.

2 The model

2.1 The logit model

Let y_i be the binary outcome for subject $i = 1, \ldots, n$, where y_i takes the value 1 if the event of interest occurred and zero otherwise. We use a logit regression model where $E(y_i)$ is linked to the linear predictor η_i by

$$E(y_i) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}.$$

For the linear predictor η_i we consider the following specification

$$\eta_i = \mu + \mathbf{x}_i\boldsymbol{\alpha}$$

where $\mathbf{x}_i$ denotes the vector of covariates and $\boldsymbol{\alpha}$ the vector of regression effects. We assume that the covariate vectors $\mathbf{x}_i, i = 1, \ldots, n$ are centered with the null vector as mean (over subjects), so that μ is a common mean for all models.

2.2 Spike and slab priors

For the effects subject to selection we use mixture priors with a spike and a slab component

$$p(\alpha_i) = (1-\omega)p_{spike}(\alpha_i) + \omega p_{slab}(\alpha_i), \quad p(\omega) \sim \mathcal{B}(a_0, b_0)$$

The spike component concentrates its mass at values close to zero whereas the slab component has its mass spread over a wide range of plausible values for the regression coefficients. We consider two different specifications: the PM (point mass) prior, where the spike is a point mass at zero, $p_{spike}(\alpha) = I_{\{0\}}(\alpha_i)$ and the slabs are independent normal or t-distributions or specified by Zellner's g-prior. In the second version both spike and slab are specified as a normal mixture of inverse Gamma (NMIG) distributions, as in Ishwaran and Rao (2005) and Konrath et. al. (2008). Introducing indicator variables $\delta_j,\ j = 1, \ldots, k$ where δ_j takes the value 1, if α_j is allocated to the slab component, the spike and slab prior can be specified hierarchically as

$$\begin{aligned} p(\delta_j = 1|\omega) &= \omega, \quad p(\omega) \sim \mathcal{B}(a_0, b_0) \\ p(\alpha_j|\delta_j) &= (1-\delta_j)p_{spike}(\alpha_j) + \delta_j p_{slab}(\alpha_j) \end{aligned}$$

and – as small effects should be assigned to the spike component – variable selection can be based on the posterior probabilities $p(\delta_j = 1|\mathbf{y})$.

3 Inference

For both prior specifications variable selection can be implemented by a MCMC sampling scheme involving the following steps:

(I) Variable selection: Sampling the vector of indicators $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_k)$

(II) Parameter estimation: Sample the vector of regression coefficients $\boldsymbol{\alpha}$ conditional on $\boldsymbol{\delta}$.

The concrete sampling steps are different, depending on the type of the prior: Whereas under the NMIG prior $\boldsymbol{\delta}$ can be drawn from the conditional

posterior $p(\boldsymbol{\delta}|\mathbf{y}, \boldsymbol{\alpha})$, for the PM prior $\boldsymbol{\delta}$ has to be drawn from the marginal likelihood $p(\boldsymbol{\delta}|\mathbf{y})$ integrating over the parameters subject to selection. In step (II) under the PM prior only coefficients with $\delta_j = 1$ have to be sampled, as those with $\delta_j = 0$ are restricted exactly to zero. As under the NMIG prior no covariates drop out of the model, the model dimension is not reduced during MCMC and the vector of all regression coefficients $\boldsymbol{\alpha}$ has to be drawn.

In contrast to the NMIG prior which allows draws from the full conditionals, the PM prior requires evaluation of marginal likelihoods in each iteration, which is computationally demanding except for normal regression models under conjugate priors. Therefore we make use of the auxiliary mixture sampler developed in Frühwirth-Schnatter and Frühwirth (2010) which leads to a representation of the logit model as a normal regression model. This auxiliary mixture sampler is based on the interpretation of a logit model as a random utilities model (RUM) and a mixture approximation of the standard logistic distribution.

4 Simulation

As variable selection is based on the posterior probabilities $p(\delta_j = 1|\mathbf{y})$, particular interest is in the efficiency of their estimates from the MCMC output. Table 1 presents the results from a simulation study where we generated $n = 200$ binary observations with 8 correlated normal regressors $\mathbf{x}_i$ where $cor(x_{ij}, x_{il}) = \rho^{|j-l|}$ for $\rho = 0.5$ and small regression effects (to avoid posterior probabilities which are close to zero or one). Variable selection was implemented for the NMIG prior using an IWLS proposal MH-algorithm as well as auxiliary mixture sampling and for the point mass prior with three different specifications of the slab component. As estimated posterior inclusion probabilities are very similar for all priors and implementations, only those obtained under the PM prior with normal slab are shown. For each implementation the inefficiency factors of the estimated posterior inclusion probabilities

$$\hat{p}_j = \hat{p}(\delta_j = 1|\mathbf{y}) = \frac{1}{M}\sum_{i=1}^{M} p(\delta_j^{(i)} = 1|\mathbf{y}, \sim)$$

are reported. Inefficiencies are higher for the NMIG prior than for the PM prior, in particular when evidence is not clear for or against inclusion of a covariate. Though the NMIG allows faster sampling than the PM prior, the resulting mean effective sample size per second (ESS/sec.) is much smaller.

5 Identifying risk factors for complications in ERCP

In our application we analyse data from an assessment project for ERCP (endoscopic retrograde cholangiopancreatography) collected in 29 Austrian

TABLE 1. Comparing spike and slab priors for simulated data (n=200, k=8) based on M = 100 000 draws after a burn-in of 40 000 draws.

		IWLS	Auxiliary mixture sampling			
α_i^{tr}	$\hat{p}(\delta_j = 1)$	NMIG	NMIG	$\mathcal{N}(0,5)$	$t_4(0,5)$	g-Prior
-0.4	**0.67**	**238.6**	**246.5**	**16.8**	**14.0**	**14.0**
-0.4	**0.52**	**198.8**	**206.2**	**16.4**	**13.7**	**14.4**
0	0.13	40.2	35.1	5.1	4.8	4.9
-0.4	**0.38**	**109.4**	**126.7**	**14.0**	**11.4**	**11.3**
-0.4	**0.92**	**194.9**	**190.3**	**14.4**	**11.4**	**10.9**
0	0.06	17.0	13.8	2.8	3.1	3.1
-0.4	**0.11**	**29.2**	**31.2**	**4.4**	**4.4**	**4.7**
0	0.06	15.8	13.7	2.6	2.7	2.7
CPU (in sec.)		361.8	339.2	872.4	807.01	767.6
mean ineff.factor		105.5	108.0	9.6	8.2	8.2
mean ESS/sec.		9.9	11.8	28.6	31.3	32.6

sites in 2006 and 2007. To identify potential patient and procedure-related factors associated with the risk of complications after ERCP, data of those 3241 patients, who experienced their first ERCP were used and 37 covariates were defined as potential risk factors. As the risk of complications is small and some risk factors are rare, separation causing a monotone likelihood occurs for 3 binary covariates. Clustering of the data within clinics and doctors is taken into account by including two types of random effects, one for the clinic and another for the endoscopist carrying out ERCP. To determine whether risk is homogeneous across clinics and doctors both methods are extended to random effects selection, yielding very similar results, with high evidence for heterogeneity among clinics, but small evidence for additional heterogeneity among the endoscopists within clinics.

References

Frühwirth-Schnatter, S. and Frühwirth, R. (2010). Data augmentation and MCMC for binary and multinomial logit models. In Kneib, T. and Tutz, G., editors, *Beitrag zur Festschrift für Ludwig Fahrmeir anlässlich seines 65. Geburtstages*, pp. 111-132, Heidelberg: Physica-Verlag.

Ishwaran, H. and Rao, S. J. (2005). Spike and slab variable selection; frequentist and Bayesian strategies. *Annals of Statistics*, 33:730–773.

Konrath, S., Kneib, T. and Fahrmeir, L. (2008). Bayesian regularisation in structured additive regression models for survival data. Technical report.

Modelling an exposure–outcome relationship accommodating potential confounders on the causal path using a latent-class model

Robert M. West[1], Mark S. Gilthorpe[1], Wendy J. Harrison[1],Amy Downing[1], David Forman[12]

[1] Centre for Epidemiology & Biostatistics, School of Medicine, University of Leeds, Leeds, LS2 9JT, UK
[2] Northern & Yorkshire Cancer Registry & Information Service, Bexley Wing, St James's Institute of Oncology, Leeds, LS9 7TF, UK

Abstract: Issues of the reversal paradox and of measurement error are addressed through the use of latent class analysis. Latent classes at two levels avoids inappropriate distributional assumptions.

Keywords: Latent-class regression; Confounder; Causal path; Measurement error.

1 Background

We model the relation between 3-year survival and socioeconomic background (SEB) amongst colorectal cancer patients whilst simultaneously considering the stage of disease at diagnosis. This is a common research question within epidemiology: to determine an exposure–outcome relationship with an additional factor introducing potential confounding. It presents analytical challenges for various reasons. In this example, many studies have shown that patient survival after diagnosis of cancer is better amongst those living in affluent areas than those in deprived areas. Consequently, such studies include stage of disease at diagnosis as a covariate to 'explain' some of the outcome differences. The difficulty is that stage lies on the causal path from socioeconomic background (SEB) to survival, through late presentation, potentially due to reduced health awareness amongst those in more deprived areas. This raises concerns regarding interpretability due to the reversal paradox, where effect estimates can be biased. Examining the relationship between SEB and survival for each stage separately does not circumnavigate this problem as it is equivalent to modelling the joint effects of SEB, stage, and their statistical interaction. A further problem is that staging data may be incomplete and missing data can also bias effect estimates, with bias exacerbated within interactions.
In 2008, NYCRIS reported 24.1% of incomplete stage data (UKACR 2008). Classification of stage is imprecise as the quality of pathology can lead to patients being classified incorrectly (Quirke and Morris 2006) or being understaged (Morris *et al.* 2007), e.g. to be diagnosed at Dukes stage C

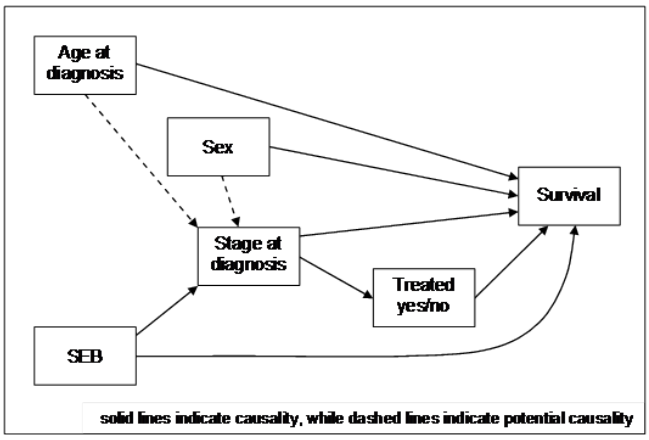


FIGURE 1. Directed acyclic graph illustrating potential casual relationships between relevant variables.

there will be nodal involvement, and the number of nodes retrieved from a resection specimen is highly variable.

We seek to alleviate these issues by using latent-class analysis (LCA) to model the data, opposed to the standard regression approach. The resulting latent classes contain differing proportions of the study population and the relationship between SEB and survival may vary across classes. Latent classes may correspond to specific patient or tumour features that permit post-hoc categorisation according to the outcome (e.g. 'good', 'moderate', and 'poor' survival). If the variable stage were to predict class membership, the resultant classes would likely exhibit variation in the proportions of individuals with early-, medium- and late-stage disease. Either way, the LCA approach facilitates the interpretation of the impact of SEB in relation to survival across patient classes akin to including stage and its interaction with SEB within the standard regression model, but without stage having to be included in the standard regression part, avoiding potential bias. LCA is undertaken in a multilevel framework to account for the clustering of patients within NHS Trusts.

LCA extended to a multilevel setting involves discrete latent variables at any level, which may be regarded as a semi-parametric approach, where the hitherto continuous latent factor at the upper level becomes categorical latent classes, replacing any Gaussian assumptions. Both patients and Trusts may thus be categorised into latent classes. Patients are not assigned randomly to Trusts and Trusts are not randomly allocated geographically. Within all health regions, Trust services have been derived historically due to many pressures operating locally: political, financial and other pragmatics issues. Therefore, standard multilevel modelling, which assumes that the variation amongst Trusts is normally distributed, may not be appro-

priate. Instead of a Trust-level continuous latent variable (assumed to be normal with mean zero and variance estimated from the data), an alternative and natural extension is to consider a discrete latent variable, where no standard distributional assumptions are required.
This paper develops the latent-class methodology of Downing *et al.* (2009) and in turn is developed further in Harrison (2010).

2 Data and Methods

Patients with colorectal cancer (ICD10 codes C18, C19 and C20) diagnosed between 1998 and 2004 and resident in the Northern and Yorkshire regions were identified from the Northern and Yorkshire Cancer Registry and Information Service (NYCRIS) database. Patient age, sex, tumour stage at diagnosis (using the Dukes classification), Trust of diagnosis and whether or not the patient received treatment were extracted. SEB was defined at the 2001 enumeration district level of residence (lower super output area) using the Townsend Index and matched to the patient using their postcode of residence. The primary outcome was the binary variable dead/alive at three years following diagnosis, which is clinically meaningful since colorectal cancer has a median survival of approximately 2–3 years. A total of 24,640 patients were available for analysis.
Patient age at diagnosis and Townsend score (SEB) were continuous measures and both exhibited a nonlinear relationship with 3-year survival, so a quadratic term for age was included and by 'trimming' SEB (assigning rare values $> \pm 5.0$ as ± 5.0) it became possible to avoid higher order terms for Townsend score. Stage was only included as a class predictor in the LCA model, rather than as a fixed-effect covariate. The number of latent classes at any level was determined by sequentially increasing from one in order to identify the optimum model according to a number of model-fit criteria. We examined BIC, AIC and changes in log likelihood (LL) and selected models that minimise all three criteria, while providing a useful model and informative results. We also observed classification error (CE) at both subject and group levels. CE is the difference between modal assignment to classes and probabilistic assignment as a proportion of the total number of patients or Trusts. A lower CE may be favoured because it could result in greater interpretability of the latent classes when considering individuals or Trusts.
In mathematical terms, the model to be fitted is of the form:

$$f(y) = \sum_{TC=1}^{nTC} P(TC) \sum_{PC=1}^{nPC} P(PC|TC) f(y|PC) \tag{1}$$

where PC indexes the patient classes, TC indexes the Trust classes, y is the outcome for the mortality status at 3 years and f is the likelihood function for the logistic regression that depends upon covariates. Here the number

Summary Stats	Class 1	Class 2	Class 3
Class Size	42.2%	30.5%	27.3%
Overall Prev	10.1%	68.9%	98.5%
Ref Group Prev	6.3%	68.8%	97.3%
Class Profiles	Class 1	Class 2	Class 3
Stage A	22.5%	6.4%	0.5%
Stage B	46.7%	18.8%	7.7 %
Stage C	27.1%	30.1%	16.2%
Stage D	0.5%	12.4%	68.6%
Missing	3.1%	32.3%	7.0%
Treatment	98.4%	75.9%	69.1%
Covariate		OR (95% CI)	
Female	0.60 (0.46–0.77)	0.84 (0.61–1.15)	1.75 (0.48–6.30)
SEB (per SD)	1.21 (1.07–1.37)	1.59 (1.31–1.92)	0.99 (0.55–1.77)
Age (per 5 yr)	2.18 (0.83–5.75)	2.53 (2.00–3.21)	0.58 (0.22–1.53)
Age^2 (per 5 yr)	1.00 (0.96–1.03)	1.01 (1.00–1.02)	1.06 (0.96–1.16)

TABLE 1. Results for the patient classes in the 3-patient-, 2-Trust-class multilevel logistic regression model: odds ratio of death within 3 years. There were 12,856 (52.2%) deaths in the entire study population; the reference group comprised males, aged 71.2 years classified as Stage I at diagnosis, and attributed a Townsend score of zero.

of patient classes nPC is constrained to be the same in all Trust classes, parameters in each patient class equal in each Trust, and there are logistic models for the allocation of latent classes at both levels. Stage enters in the patient class allocation model and other covariates only through the conditional likelihood.

3 Results

In the study population, 12,856 patients (52.2%) died within three years. The preferred multilevel latent class model exhibited three patient classes and two Trust classes. Patients were assigned to either a large good-prognosis group (PC1), a small reasonable-prognosis group (PC2), or an even smaller poor-prognosis group (PC3).

The stage profile differs across patient classes. PC1 (good prognosis) corresponds to early stage diagnosis, with 69% of the stage A/B patients versus 28% of the stage C/D patients. PC2 (reasonable prognosis) corresponds to more evenly balanced staging at diagnosis, with 25% of the stage A/B patients and 43% of the stage C/D patients; this class also contained 32% of patients with missing stage. PC3 (poor prognosis) corresponds to late stage diagnosis, with only 8% of the stage A/B patients versus 85% of the

Model Statistics	T Class 1	T Class 2
Class Size	69.7%	30.3%
Overall Mortality	52.8%	51.0%
Subject Class Profiles	T Class 1	T Class 2
Good prognosis group	41.5%	43.8%
Reasonable prognosis group	29.2%	33.4%
Poor prognosis group	29.3%	22.9%

TABLE 2. Results for the Trust classes in the 3-patient-, 2-Trust-class multilevel logistic regression model.

stage C/D patients. Of particular note is the proportion of stage D patients in each class, with only 0.5% in PC1, 12% in PC2, and 69% in PC3.
The impact of deprivation differs across the patient classes. In PC1 and PC2, living in a more deprived area was clearly associated with increased odds of death (PC1 OR=1.21, 95% CI=1.07 to 1.37; PC2 OR=1.59, 95% CI=1.31 to 1.92 per SD increase in Townsend score). In PC3, the association was less clear (OR=0.99, 95% CI=0.55 to 1.77). This suggests that the potential impact of SEB on 3-year survival from colorectal cancer may operate somewhat differently for differently staged individuals, with SEB having less impact for those with later-staged disease.
At the Trust level, partly due to aggregation, differences are more subtle. For this formulation of the model, the Trust effect is regarded as a nuisance, the primary focus being the assessment of deprivation on patients with appropriate adjustment for 'stage' and allowing for clustering within Trusts. The indication here however is that the Trust effects are small compared with the effect of patient classes.

4 Discussion

By removing stage from the fixed-part of the regression model, we have avoided the reversal paradox and also minimised bias due to measurement error or incomplete data. As resultant patient classes correspond well to stage at diagnosis, we have determined how SEB impacts upon 3-year survival across the different stage groupings, without introducing a statistical interaction term, which would further exacerbate potential bias due to measurement error or incomplete data. With stage included as a class predictor, however, bias might not be entirely eradicated, since patient classes then depend, in part, upon stage, and interpretation of the impact of SEB on survival once again becomes dependent upon stage, albeit less directly than including stage as a fixed-effect covariate in the model. Since patient classes may be derived without stage as a class predictor and the same differentiation of patient subgroups across patient classes is observed akin to that

of stratification by stage, the preferred model might be that without stage used explicitly in the model process at all.
It could be more informative to explore survival as a continuous measure, e.g. using Cox's proportional hazards regression. In addition, SEB is measured at the area-level and so should be considered as a separate level, effectively cross-classified with the Trust level. Both model improvements are theoretically possible and implementable in due course, following software developments. This study nevertheless illustrates the principles of using LCA to avoid some of the challenging problems in epidemiology of interpreting the exposure–outcome relationship whilst considering an additional factor that might present some analytical challenges were it considered in the standard regression model.

5 Conclusion

For those patients presenting with early- or mid-stage disease, their characteristics may help determine their chances of dying from colorectal cancer, whilst for patients presenting with late-stage disease, their characteristics are less likely to affect mortality. LCA models have utility in circumnavigating the reversal paradox, i.e. when considering potential confounding factors that lie on the path between exposure and outcome.

References

Downing, A., Harrison, W.J., West, R.M., Forman, D., and Gilthorpe, M.S. (2009) Latent class modelling of the association between socioeconomic background and breast cancer survival status at 5 years whilst incorporating stage of disease. *Journal of Epidemiology and Community Health*, in press (epub), PMID: 19692736.

Harrison W.J., West, R.M., Downing, A., Forman, D., and Gilthorpe, M.S. (2010) Multilevel latent-class modelling of patient casemix. *IWSM, Glasgow*.

Morris, E.J.A., Maughan, N.J., Forman, D., and Quirke, P. (2007) Identifying stage III colorectal cancer patients: the influence of the patient, surgeon and pathologist. *Journal of Clinical Oncology*, **25**, 2573–2579.

Quirke, P., and Morris, E.J.A. (2006). Reporting colorectal cancer. *Histopathology*, **50**, 103–112.

UKACR (2008) Quality and Performance Indicators 2008: Final [online]. [cited 2-2-2009]; Available from:

`http://82.110.76.19/quality/UKACR%20report2008_final.pdf`.

Zero Augmentation: A method for fitting zero-modified count models that allows both zero-inflation and zero-deflation

Paul Wilson[1]

[1] School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway.

Paul.Wilson@nuigalway.ie

Abstract: The concept of zero-inflation is now well established, and several software packages exist that fit zero-inflated models. Whilst there is some mention of zero-*deflation* in the literature, there is very little published evidence of research into zero-*modified* models that allow for both zero-inflation and zero-deflation within the one model. We present a very simple method that enables the fitting of zero-modified models via *any* software that fits zero-inflated models, and investigate the benefits of fitting zero-modified models.

Keywords: Zero-modification, Zero-inflation, Zero-deflation

1 Introduction

Whilst the concept of zero-inflation is now well established, and several software packages exist (e.g. the R-packages PSCL, ZIGP, VGAM and GAMLSS) that fit zero-inflated models, i.e. models of the form:

$$f(y;\Theta) = \begin{cases} \gamma + (1-\gamma)f(0;\Theta) & y = 0 \\ \\ (1-\gamma)f(y;\Theta) & y = 1,2,3,\ldots \end{cases} \tag{1}$$

where $\gamma > 0$, there is a near absence of work concerning zero–*modified* models that allow for both zero-inflation and zero-deflation within the same dataset, and thus permit negative (as well as positive) values of γ. Zero-deflation may arise as a consequence of under-reporting of zero counts. For example, say a study was concerned with the distribution of the number of eggs laid in bird's nest dependent on various covariates. Certain species of birds, for instance the Whooping Crane (Grus Americana), make nests by making shallow depressions in marshy ground. Clearly it may be difficult to distinguish a natural shallow depression in the ground from one made by a whooping crane as a nest, but in which no eggs were laid. Over-classification of empty nests as natural depressions will lead to under-reporting of the

number of zero counts of the number of eggs laid in such nests, and hence zero-deflation, whereas over-classification will lead to zero-inflation.
Even if theoretically zero-deflation does not make sense in the context of the data being analysed, zero deflation may occur in the observed data for certain combinations of covariates. If a model that is constrained to return positive values of γ is fitted to observed data that is zero-deflated for certain combinations of covariate values, a model fitting mechanism that employs an EM algorithm to estimate the proportion of the data that arises from a perfect-zero distribution will return a value of zero, or possibly a small positive value. This will both influence the estimates of positive values of γ for other covariate combinations, and, perhaps more importantly will affect the estimation of the mean: to compensate for the increase to zero of a negative γ estimate a greater value of the estimate of the mean will occur.
Whilst a Bayesian approach to zero-modified models has been proposed by Angers and Biswas (2003), the only frequentist approach would appear to be that of Dietz and Böhning (2000). Their method incorporates the unusual link-function:

$$\eta = \eta(\gamma) = \log\left(\frac{1-\gamma}{e^{\mu}/(e^{u}-1)-(1-\gamma)}\right) \tag{2}$$

where $\gamma = 1 - \frac{e^{X\beta}}{1+e^{X\beta}}\frac{e^{\mu}}{e^{\mu}-1}$
for the zero-modification parameter. This technique has the undesirable property that the link function varies according to the Poisson mean, and is not readily implementable using standard software packages for fitting zero-inflated models. The method of zero-augmentation introduced here enables *any* zero-modified models to be fitted using *any* software, and furthermore, *any* link functions available for fitting such models may be utilised.

2 Zero Augmentation

We define *Zero-Augmentation* of a dataset to be the artificial addition of zeros to the response variable of that dataset. The proposed system of zero-augmentation is to replicate the data $\kappa - 1$ times to form a "κ-augmented dataset", whilst the values of the various covariates remain the same in the replicated data, the values of the response variable are replaced by zeros. For instance if the original data (where the response variable is in the final column) is:

1	0	2	1	0
2	1	3	1	5
3	1	1	2	3

then the 2–augmented data is:

1	0	2	1	0
2	1	3	1	5
3	1	1	2	3
1	0	2	1	0
2	1	3	1	0
3	1	1	2	0

The idea behind zero-augmentation is extremely simple: the zeros added to the response variable by zero-augmentation are by definition extra-zeros. If the model fitting mechanism were to classify all such augmented zeros as extra-zeros, then zero-inflation parameter estimates for the original, non-augmented data may be deduced from those of the augmented data, (for example, for 2-augmented data a parameter value of $\gamma^\star$ for the zero-modification parameter may be shown to correspond to an estimate of

$$\hat{\gamma} = 2\gamma^\star - 1 \tag{3}$$

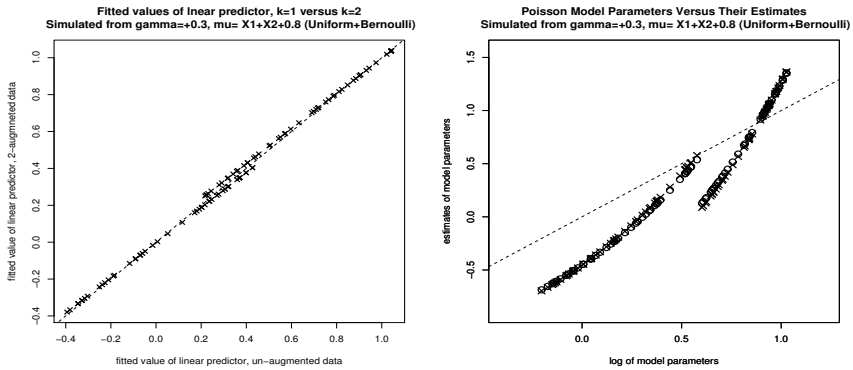
after 2-augmentation). Hence a positive γ parameter in the data of, say, 0.3 will return an estimate of approximately 0.65 when the model is fitted to the 2-augmented data which may then be "converted back" to an estimate of approximately $(2 \times 0.65 - 1) = 0.3$ for the original data. However if fitting the model to the 2-augmented data returns an estimate of $\gamma^\star = 0.2$, this corresponds to an estimate of $\hat{\gamma} = (2 \times 0.2) - 1 = 0.6$ for the original data, indicating zero-deflation. Other model parameter estimates should remain unchanged by zero-augmentation, the structure of the non–extra zeros count data has not been altered. In practice the model fitting procedure may not classify all augmented zeros as extra zeros, but any resulting bias is not serious. In the vast majority of cases, 2-augmentation is sufficient to detect and model any zero-deflation that exists in the data, for the remainder of this paper we present examples of the technique restricting ourselves to non-augmentation and 2-augmentation.

2.1 Example 1

To illustrate that the estimates of the linear predictors of the mean are extremely similar under non-augmentation and 2-augmentation, 100 data were simulated from a zero-inflated Poisson model where $\gamma \sim 0.3$ and $\mu \sim X_1 + X_2 + 0.8$, where $X_1 \sim U(0,1)$ and $X_2 \sim bernoulli(0,1)$, and zero-inflated models fitted (using log and cloglog links). The left hand diagram of Figure 1 plots the values of the estimated values of the linear predictors of the Poisson means against each other, whilst the right-hand diagram plots the values of the estimates under non-augmentation ($\times$) and 2-augmentation ($\circ$) against the parameters of the model from which the data was simulated. We see that these estimates are very similar. The

values of $\hat{\gamma}$ obtained from fitting the model to the original data and the 2-augmented data (after applying formula of equation (3) was 0.29 in both cases. The log-likelihood of the models fitted to the data calculated using the parameter estimates from the non-augmented and augmented data respectively were almost identical at -134.63 and -134.59 respectively.

FIGURE 1. Comparative fits of the Poisson parameters of zero-modified Poisson data under non-augmentation and 2-augmentation



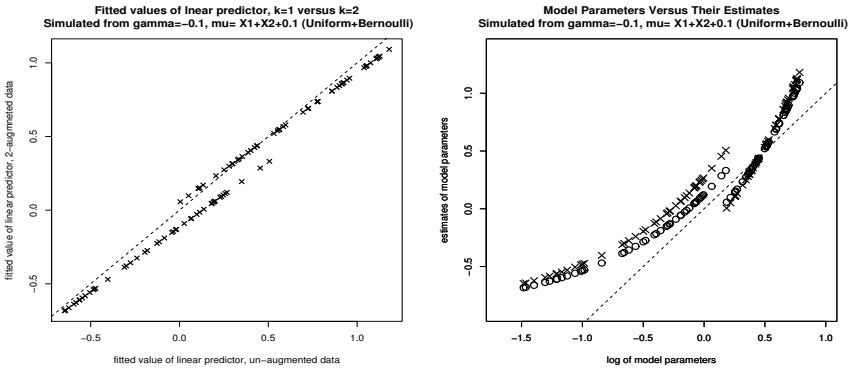
2.2 Example 2

The diagrams of Figure 2 pertain to 100 data simulated from zero-deflated Poisson data where $\gamma \sim -0.1$ and $\mu \sim X_1 + X_2 + 0.1$, where $X_1 \sim U(0,1)$ and $X_2 \sim bernoulli(0,1)$. We see from the left hand diagram that here the estimates are not similar, indicating zero-deflation. When the zero-inflated model was fitted to the non-augmented data a value of zero was returned for $\hat{\gamma}$, whereas, after using the formula of equation (3) an estimate of $\hat{\gamma} = -0.08$ is returned when the model is fitted using the 2-augmented data. Note that, as a consequence of this more accurate estimation of γ, in general the estimates of the values of the Poisson means ($\circ$) are closer to the values of the means of the model from which the data was simulated. Here the log-likelihood of the data calculated using the parameters estimated from the augmented data was -133.67, whereas that obtained from the non-augmented data was slightly poorer at -134.40

2.3 Example 3

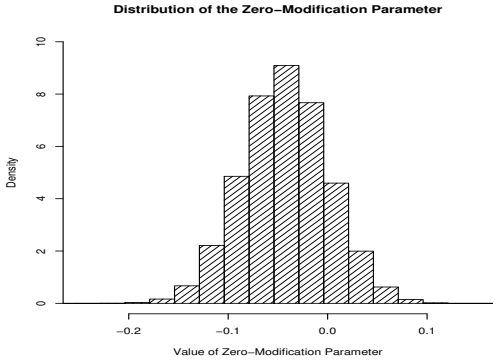
For this example, 300 data were simulated from the zero-modified Poisson model $\gamma \sim \frac{\mu-3}{10}$ where $\mu \sim 0.3X_1 + 0.2X_2 + 0.5X_3 + 0.3X_4 + 0.5$, and $X_1 \sim N(0,1)$, $X_2 \sim U(0.5,2)$, $X_3 \sim N(3,0.5)$ and $X_4 \sim U(0.2,2)$. Figure

FIGURE 2. Comparative fits of the Poisson parameters of zero-modified Poisson data under non-augmentation and 2-augmentation



3 illustrates the distribution of the resultant zero-modification parameters. As is apparent both zero-inflation and deflation are present in the model, the values of the zero inflation parameter being approximately normally distributed with mean -0.042 and standard deviation 0.042.

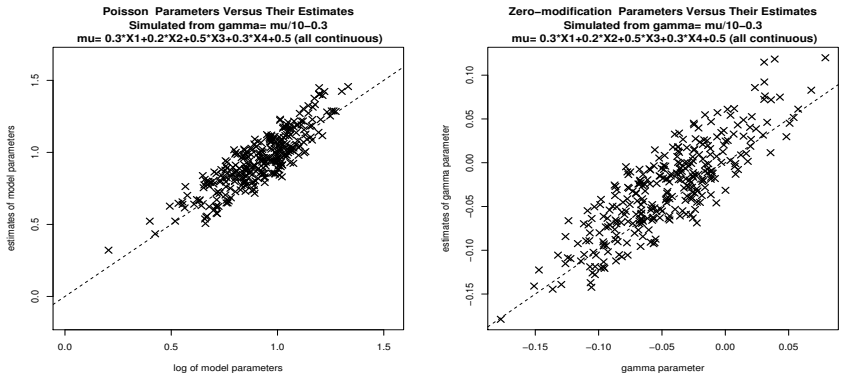
FIGURE 3. Distribution of the Zero-Modification Parameter, Example 3



When it was attempted to fit a zero-inflated model to the non-augmented data, (using logit and log links), the fitting algorithm failed due to the considerable zero-deflation. The fitting of the 2-augmented data was successful however. The left hand diagram of Figure 4 plots the estimates of the linear predictors of the Poisson means against the parameters of the model from which the data was simulated. We see that this diagram indicates that the fitted means correspond to what is expected from data simulated

from such a model. The right hand diagram plots the estimates of the fitted values of the of the zero-inflation parameters (adjusted using the formula of equation (3)) against the parameters of the model from which the data was simulated. Again, these indicate a good degree of fit.

FIGURE 4. Fits of the parameters of zero-modified Poisson data under 2-augmentation



3 Conclusion

Zero-augmentation is a very simple technique that allows data that contains both zero-inflation and zero-deflation to be modelled using standard software. Whilst the discussion above concentrates on the estimation of the means of zero-modified Poisson data, the technique may be utilised to estimate any parameter of any zero-modified model. The development of this technique is in the early stages, future research areas include quantification of the amount if bias introduced by the augmentation procedure, the refinement of the technique to compensate for such bias and the development of associated tests to determine whether zero-deflation may be present in the data being analysed.

References

Angers JF, Biswas A (2003) A Bayesian analysis of zero-inflated generalized Poisson model. *Computational Statistics and data Analysis* Vol. 42, pp. 37–46

Dietz D, Böhning D (2000) On the estimation of the Poisson parameter in zero-modified Poisson models. *Computational Statistics and data Analysis* Vol. 34, pp. 441–459

Modelling by using a smooth family of empirical distributions and an application to failure time data

Bruce J. Worton[1]

[1] Bruce.Worton@ed.ac.uk - School of Mathematics and Maxwell Institute for Mathematical Sciences, The University of Edinburgh, James Clerk Maxwell Building, King's Buildings, Mayfield Road, Edinburgh EH9 3JZ, U.K.

Abstract: This paper considers the construction of a smooth nonparametric family of empirical distributions index by a specified parameter of interest and shows its usefulness in statistical inference. An application to failure time data illustrates how the family of distributions may be employed to calculate a likelihood for the parameter of interest.

Keywords: Bootstrap; Likelihood; Tilting; Smoothing populations.

1 Introduction

A nested bootstrap method may be used to generate a nonparametric family of empirical distributions index by a specified parameter of interest. Suppose we have an estimator T for a parameter θ and a data set $x_1, \ldots, x_n$, which is assumed to be a random sample from a distribution function F. We may generate a first-level bootstrap sample $x_1^*, \ldots, x_n^*$ from $\widehat{F}$, the empirical distribution of $x_1, \ldots, x_n$, and calculate the estimate t^* associated with the resample $x_1^*, \ldots, x_n^*$. We consider such a first-level bootstrap sample as a population $\mathcal{P}^*$ with parameter value t^*. Repeating this step M times produces populations $\mathcal{P}_1^*, \ldots, \mathcal{P}_M^*$ with parameter values $t_1^*, \ldots, t_M^*$. The populations may be used like parametric distributions, but some care is needed due to the erratic variability resulting from the resampling.
Davison, Hinkley and Worton (1992) used the above approach to define an analogue of partial likelihood as follows. For each $\mathcal{P}_m^*$ and t_m^*, $m = 1, \ldots, M$, use a second-level of bootstrapping to estimate the density of T^{**}, the estimator computed from a second-level resample $x_1^{**}, \ldots, x_n^{**}$. Evaluating each density at t, the observed value of T for the original data set, gives M likelihood *points* at parameter values $t_1^*, \ldots, t_M^*$. A smooth likelihood may be obtained by curve fitting, but has the disadvantage that it does not correspond directly to an underlying family of distributions. In this paper we derive an efficient nested bootstrap approach to obtain a family of distributions that resembles a parametric family of distributions.

2 Methodology

A convenient way to smoothly aggregate samples for a target value of the parameter $\theta = \theta^0$ is to average the populations $\mathcal{P}_1^*, \ldots, \mathcal{P}_M^*$ which have been generated to have an appropriate density of their parameters $t_1^*, \ldots, t_M^*$ by using the exponential tilted resampling defined below, with a kernel-type smoother (Davison, Hinkley and Worton 1995)

$$p_i^*(\theta^0, \epsilon) \;\propto\; \sum_{m=1}^{M} w\left(\frac{\theta^0 - t_m^*}{\epsilon}\right) p_{mi}^*, \qquad i = 1, \ldots, n,$$

where $p_{m1}^*, \ldots, p_{mn}^*$ denote the relative frequencies of $x_1, \ldots, x_n$ for population $\mathcal{P}_m^*$, $m = 1, \ldots, M$, with a chosen bandwidth $\epsilon > 0$ and normal kernel function $w(.)$. Canty, Davison, Hinkley and Ventura (2006) have successfully used this in the context of bootstrap diagnostics, but there are still only fairly limited guidelines on how much to smooth.

Exponential tilted resampling (Hinkley and Shi 1989) can be used as an efficient method to increase the generation of values of t^* in regions of high log bootstrap likelihood variability. The resampling probabilities of the data are taken as $\Pr(X^* = x_i) \propto \exp(\alpha I_i)$, for a specified tilting constant α, to give $T^* \sim \mathrm{N}(t + \alpha s^2, n^{-1}s^2)$. This may be done most efficiently if we generate first level resamples proportional to the conditional variance of log bootstrap likelihood. Other methods may be extremely inefficient.

An approximate expression for the conditional variance of log bootstrap likelihood based on the sample mean estimator, $T = n^{-1}\sum_{i=1}^{n} X_i$, of the population mean, $\theta = E(X) = \int x dF(x)$, can be obtained using the properties of the multinomial distribution. It can be shown that it is given by

$$v(\theta) = \frac{\partial l(t^*)}{\partial \mathbf{r}^*}^T \mathbf{C}_\theta \frac{\partial l(t^*)}{\partial \mathbf{r}^*}, \tag{1}$$

where $\mathbf{r}^*$ is the relative frequencies of resamples at the data values, $l(t^*)$ is an approximate expression for the log likelihood and $\mathbf{m}_\theta$,$\mathbf{C}_\theta$ are obtained from conditional mean, covariance-matrix, with $\mathbf{r}^*$ evaluated at $\mathbf{m}_\theta$. Given that $t^* = \theta$ in $l(t^*)$, it may be shown that the ith element of $\partial l(t^*)/\partial \mathbf{r}^*$, at $\mathbf{r}^* = \mathbf{m}_\theta$, is given by

$$\frac{n(x_n - x_i)(\theta^2 - x_i x_n)\{(t-\theta)^2 - n^{-1}s_\theta^2\}}{2x_n s_\theta^4},$$

where $s_\theta^2 = \sum_{i=1}^{n} \mu_i (x_i - \theta)^2$, with $(\mu_1, \ldots, \mu_{n-1})^T = \mathbf{m}_\theta$, and μ_n determined by $\sum_{i=1}^{n} \mu_i x_i = \theta$, $x_n \neq 0$.

If K samples with similar values of $t^* \approx \theta_0$ are aggregated to produce an 'average' smoothed population with parameter value $t^* \approx \theta_0$ then it turns out that $v(\theta_0)$ is reduced by a factor of K. The variance is very far from being constant as t^* varies, so we can vary K with t^*.

3 Application to modelling failure time data

We consider the construction of a smooth family of empirical distributions for the aircraft air-conditioning data from Cox and Snell (1981, aircraft 7 of Example T) of $n = 24$ time intervals in hours between repairs and failures. The family is then used to calculate a bootstrap likelihood, as plots of (unsmoothed) first-level $\mathcal{P}_m^*$ populations for the aircraft failure time data are seen to vary greatly, even for very similar t^* parameter values. The precise features of the tilting will depend nonparametrically on the characteristics of the first-level resamples.

Taking the parameter of interest θ as the mean failure time, the top panel of Figure 1 shows the approximate variance curve $v(\theta)$ for this problem which was used to obtain an efficient method for constructing populations. The variability is extremely high for the lower values of the parameter, and thus this region requires more aggregation of samples than the central parameter region near $t = 64.125$, or even the upper region, due to the nature of the failure time data.

The log bootstrap likelihood points $(t^*, l(t^*))$ using exponential tilting to generate populations, as well as the corresponding likelihood points, are also shown in Figure 1 for the unsmoothed populations, with $M = 1000$. Note the high levels of variability of log bootstrap likelihood for the lower values of θ as predicted by the top panel.

Figure 2 shows eight independent log bootstrap likelihood curves, obtained by repeat applications of the method but with the smoothed populations replacing the unsmoothed populations. Each curve was calculated by aggregating $M = 1000$ unsmoothed populations with $\epsilon = 0.3sn^{-\frac{1}{2}}$. Evidently, there is a dramatic reduction in the variability relative to the corresponding plot in Figure 1. This method provides a much more effective use of first-level bootstrap samples than a more basic method, and it seems more desirable for likelihood to vary smoothly over an underlying family.

4 Discussion and conclusions

We have shown that by using the properties of first-level bootstrap samples it is possible to define a smooth family of nonparametric distributions indexed by a parameter of interest. We may contrast this approach with employing a parametric family of (empirical) distributions which varies smoothly with the parameter of interest, e.g. an empirical exponential family model. However, we conclude with the observation that modelling by using the smoothed populations seems to be an attractive compromise between (i) using the unsmoothed highly variable bootstrap populations directly and (ii) using a parametric family of empirical distributions which varies slowly with the parameter value. The proposed modelling approach has the advantage of not placing possibly unreasonable constraints on the family of distributions which may be difficult to assess in practice.

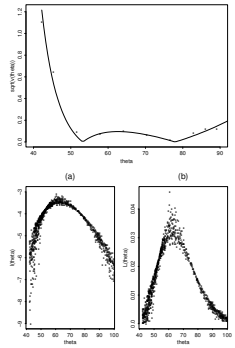


FIGURE 1. Top panel: Approximate conditional root-variance of log likelihood; Bottom panel: (a) Log bootstrap likelihood points for the mean failure time data using unsmoothed first-level populations, and (b) the corresponding bootstrap likelihood points.

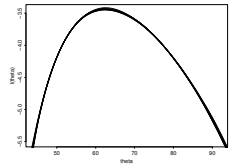


FIGURE 2. Eight independent log bootstrap likelihood curves for the mean failure time data using smoothed first-level populations.

Acknowledgments: I would like to thank Professor D.V. Hinkley and Professor A.C. Davison for helpful discussions.

References

Canty, A.J., Davison, A.C., Hinkley, D.V., and Ventura, V. (2006). Bootstrap diagnostics and remedies. *Canadian Journal of Statistics*, **34**, 5-27.

Cox, D.R., and Snell, E.J. (1981). *Applied Statistics: Principles and Examples*. London: Chapman & Hall.

Davison, A.C., Hinkley, D.V., and Worton, B.J. (1992). Bootstrap likelihoods. *Biometrika*, **79**, 113-130.

Davison, A.C., Hinkley, D.V., and Worton, B.J. (1995). Accurate and efficient construction of bootstrap likelihoods. *Statistics and Computing*, **5**, 257-264.

Hinkley, D.V., and Shi S. (1989). Importance sampling and the nested bootstrap. *Biometrika*, **76**, 435-446.

Modelling covariance structures for multivariate longitudinal data

Jing Xu[1], Gilbert MacKenzie[1]

[1] Centre of Biostatistics, University of Limerick, Ireland. ENSAI, Rennes, France
Email: jing.xu@ul.ie; gilbert.mackenzie@ul.ie

Abstract: The analysis of multivariate longitudinal data can be challenging because of the existence of correlations between responses which may be caused by multiple variables and repeated measurements. Therefore, one major task in analyzing these data is to efficiently model the covariance matrices $cov(y_i) = \Sigma_i$ for $i = 1, \cdots, n$ subjects. In this paper, we developed a data-driven method to model the covariance structures. By this method, constrained and hard-to-model parameters of Σ_i are traded in for the unconstrained and interpretable parameters. Estimates of these parameters, together with the parameters in the mean, are obtained by maximum likelihood approach, and the large-sample asymptotic properties are derived. A simulation is carried out to illustrate the method introduced.

Keywords: multivariate longitudinal data; marginal models; covariance modelling; block triangular factorization; matrix logarithm

1 Introduction

In many epidemiological studies and clinical trials, subjects are measured on several occasions with regard to a collection of response variables. Analysis of such multivariate longitudinal data involves modelling the joint evolution of the response variables over time. Consider, as an example, a study of anaemia in pregnancy (McMullan, et al 2003) carried out in Belfast. 264 patients had three visits to the clinic. For them, two blood measurements, erythropoietin Epo and haemoglobin Hb, were taken through out pregnancy. There are many similar examples (Chapman et al 2003, Thiebaut et al 2002, Newsom 2002).

However, the analysis of such a multivariate longitudinal data is complicated by a) the correlation between the responses at each time point, b) the correlation within separate responses over time, and c) the cross-correlation between different responses at different times. Therefore, one major task in analyzing these data is to efficiently model the covariance matrices $cov(y_i) = \Sigma_i$ for $i = 1, \cdots, n$ subjects. Several approaches have been developed: doubly multivariate models(DMM) analysis (Timm, 1980),

multivariate repeated measurement models with a Kronecker product covariance structure (Galecki, 1994), multivariate mixed models (Jones 1993) and structural equation modelling apparach (Hatcher, 1998). In this paper, we developed a data-driven method to model the covariance structures. We extend the idea of covariance modelling (Pourahmadi 1999) for the traditional univariate longitudinal data to the multivariate case by using block triangular factorization of Σ_i. This new method maintains most of the nice properties owned by the univariate case, i.e, the decomposition is unique, positive definiteness of Σ_i is guaranteed, the new parameters are unconstrained and have useful statistical interpretations.

2 Covariance modelling

For simplicity, the method is presented for the biovariate case in the rest of paper, although it can be straight-forwardly applied to the multivariate case.

2.1 Block triangular factorization of Σ

Let $y_{ij} = (y_{ij}^{(1)}, y_{ij}^{(2)})'$ present the observations of two response variables for the i-th individual at j-th time point ($i = 1, \cdots, n; j = 1, \cdots, m$). Further let $y_i = (y_{i1}', \cdots, y_{im}')'$. Denote the covariance matrix of y_i by Σ. Here we assume that the covariance matrices of y_i are homogeneous across subjects. Noting that Σ is positive definite, Σ can be factorized block-triangularly as (see Hamilton 1994)

$$T\Sigma T' = D, \quad \text{or} \quad \Sigma^{-1} = T'D^{-1}T,$$

where T is a block lower triangular with 2×2 identity matrices as diagonal entries and D is a block-diagonal matrix with positive definite 2×2 matrices as diagonal entries. It is easily seen that Σ is positive definite if and only if D is positive definite and the decomposition is unique and has the following statistical interpretation: the block matrices, denoted by $\Theta_{i,j}$, as the below-diagonal entries of T are the negatives of the coefficient matrices of $\widehat{y}_{ij} = \mu_{ij} + \sum_{k=1}^{j-1} \Theta_{j,k}(y_{ik} - \mu_{ik})$, the linear least-squares predictor of y_{ij} based on its predecessors $y_{ij-1}, \cdots, y_{i1}$, and the block diagonal entries, denoted by D_j, of D are the prediction error covariances $D_j = cov(y_{ij} - \widehat{y}_{ij})$, for $1 \leq j \leq m$. With this decomposition, the $\frac{1}{2}2m(2m+1)$ constrained and hard-to-model parameters of Σ can be traded in for the $\frac{1}{2}2m(2m+1)$ unconstrained and interpretable parameters $\Theta_{j,k}$, $\log D_j$ (see subsection 2.2) for $1 \leq j \leq m$ and $1 \leq k \leq j-1$. We refer to the new parameters $\Theta_{j,k}$'s and D_j's as the autoregressive coefficient matrices and the innovation covariance matrices of Σ.

2.2 Matrix logarithm of D

Since D is positive definite, i.e, all the diagonal entries $D_1, \cdots, D_m$ are positive definite, the matrix logarithm of D_j can now be defined by

$$A_j = \log D_j$$

basing on the spectral decomposition of D_j. The positive definiteness of D_j is guaranteed by the definition of matrix exponential(Chiu et al. 1996).

2.3 Linear covariance models

Since $\Theta_{j,k}$ and $\log D_j$ are unconstrained, we may model them in terms of covariates. For example, the polynomials of time and lag. The new parameters in the linear models for $\Theta_{j,k}$ and $\log D_j$ are denoted by the unknown vectors γ and λ.

3 Maximum Likelihood Estimation

3.1 Estimation of parameters

In the marginal linear regression models with normal distributed responses, the estimates of the parameters γ and λ in the covariance matrices, together with the parameters, denoted by β, in the mean part, can be obtained by a maximum likelihood approach. The log-likelihood of β, γ and λ, given $y_1, \cdots, y_n$, satisfies

$$2\log \ell(\beta, \gamma, \lambda | y_1, \cdots, y_n) = -mn\log(2\pi) - n\log|D| - \sum_{i=1}^{n} r_i' T' D^{-1} T r_i, \quad (1)$$

where $r_i = y_i - X_i\beta$ and X_i is the design matrix in the regression model. Fixing γ and λ in (1) creates the weighted least squares solution of β is

$$\widetilde{\beta} = \{\sum_{i=1}^{n} X_i' \Sigma^{-1} X_i\}^{-1} \sum_{i=1}^{n} X_i' \Sigma^{-1} y_i. \quad (2)$$

Secondly, given β and λ, the solution of the first derivative of γ is

$$\widetilde{\gamma} = \{\sum_{i=1}^{n} Z_i^{*'} D^{-1} Z_i^*\}^{-1} \sum_{i=1}^{n} Z_i^{*'} D^{-1} r_i, \quad (3)$$

where $Z_i^* = (r_{i1}^*, \cdots, r_{iq}^*)$ with $r_{il}^* = U_l^* r_i$. Here $U_l^* (l = 1, \cdots, q)$ are the block lower triangular matrices with off-diagonal matrices $U_{jkl} (k < j, j = 2, \cdots, m)$ and zero matrices as diagonal entries.
Denote the $(2j-1)$-th and $2j$-th elements of the vector Tr_i by the 2×1 vector $e_{ij} = r_{ij} - \widehat{r}_{ij}$ with $\widehat{r}_{ij} = \sum_{k=1}^{j-1} \Theta_{j,k} r_{ik}$ for $j = 1, \cdots, m$. By the

definition of matrix exponential and logarithm and having that $\log|D_j| = tr(A_j) = \sum_{l=1}^{d} \lambda_l tr(V_{jl})$ for $j = 1, \cdots, m$, the log-likelihood function excluding the constant becomes

$$2\log \ell(\beta, \gamma, \lambda | y_1, \cdots, y_n) \sim -mn \sum_{l=1}^{d} \lambda_l tr(\overline{V}_{\cdot l}) - n \sum_{j=1}^{m} tr\{B_j \exp(-A_j)\}, \quad (4)$$

where $\overline{V}_{\cdot l} = \sum_{j=1}^{m} V_{jl}/m$ and $B_j = \sum_{i=1}^{n} e_{ij} e_{ij}'/n$. The first and second order of derivatives of $\log \ell$ with respect to λ are derived by applying the directional derivative of the matrix exponential (Bellman 1970) to the Taylor series expansion of function (4) with respect to λ. Fixed β and γ, the solution of the estimation equation for λ can be obtained by the Newton-Raphson iterations. we denote it by $\widetilde{\lambda}$.
The iterative procedure proceeds within (2), (3) and (4) by initializing at $\Sigma = I_m$ where I_m is a $m \times m$ identity matrix and iterating until convergence to obtain the ML estimator $(\widehat{\beta}', \widehat{\gamma}', \widehat{\lambda}')'$ simultaneously.

3.2 Asymptotic properties

Briefly speaking, under some necessary regularity conditions, the ML estimators $\widehat{\theta} = (\widehat{\beta}', \widehat{\gamma}', \widehat{\lambda}')'$ is strongly consistent for the true value $\theta_0 = (\beta_0', \gamma_0', \lambda_0')'$ and the ML estimator $\widehat{\theta}$ has an asymptotically normal distribution.

4 Simulation

A small-scale simulation study in this section is conducted to investigate the adequacy of the asymptotic results.
A sample of n two-dimensional observations vectors Y_i, $i = 1, \cdots, n$ are generated normally at $m = 11$ time-points. The mean of Y_{ij} in $Y_i = (Y_{i1}, \cdots, Y_{im})$ follows

$$\mu_{ij} = \begin{pmatrix} \beta_1 + \beta_2 t_j \\ \beta_3 + \beta_4 t_j \end{pmatrix}, \quad t_j = -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5,$$

so the design matrix X_{ij} for the subject i at time-point j is given by

$$X_{ij} = \begin{pmatrix} 1 & t_j & 0 & 0 \\ 0 & 0 & 1 & t_j \end{pmatrix}.$$

The covariance matrix Σ_i of Y_i is constructed by $\Sigma_i = T^{-1} D (T')^{-1}$. The block lower triangular T has 2×2 identity matrices as diagonal entries and the negative of matrices

$$\Phi_{jk} = \begin{pmatrix} \gamma_1 + \gamma_2(t_j - t_k) & \gamma_5 + \gamma_6(t_j - t_k) \\ \gamma_7 + \gamma_8(t_j - t_k) & \gamma_3 + \gamma_4(t_j - t_k) \end{pmatrix}, \quad j = 2, \cdots, m; k < j, \quad (5)$$

as off-diagonal entries. The block-diagonal matrix D has positive definite 2×2 matrices $D_j = \exp(A_j)$, $j = 1, \cdots, m$ as diagonal matrices with A_j being

$$A_j = \begin{pmatrix} \lambda_1 + \lambda_2 t_j & \lambda_5 + \lambda_6 t_j \\ \lambda_5 + \lambda_6 t_j & \lambda_3 + \lambda_4 t_j \end{pmatrix}. \tag{6}$$

It is easy to see that the matrix (5) can be rearranged in the form $\Phi_{jk} = \sum_{l=1}^{8} \gamma_l U_{jkl}$ with $U_{jk1} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right), U_{jk2} = \left(\begin{smallmatrix} j-k & 0 \\ 0 & 0 \end{smallmatrix}\right), U_{jk3} = \left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right), U_{jk4} = \left(\begin{smallmatrix} 0 & 0 \\ 0 & j-k \end{smallmatrix}\right)$, $U_{jk5} = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right), U_{jk6} = \left(\begin{smallmatrix} 0 & j-k \\ 0 & 0 \end{smallmatrix}\right), U_{jk7} = \left(\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right)$ and $U_{jk8} = \left(\begin{smallmatrix} 0 & 0 \\ j-k & 0 \end{smallmatrix}\right)$. The similar rearrangement can be done to A_j in (6). Noting that the $(1,1)$-th 2×2 sub-matrix in the covariance matrix $\Sigma_i = T^{-1} D (T')^{-1}$ is D_1 and every element of $\log(D_1)$ is a linear function of t_1, it is easy to see that $\Sigma(\alpha) \neq \Sigma(\alpha_0)$ given that $\lambda \neq \lambda_0$ for at least one value of time point t_1 since a new starting observation time can be chosen for t_1 without affecting analysis of the data. Generally, Condition I can be always satisfied provided that the elements of $\log D_1$ are linear regression models of λ.

Table 1 summarizes the simulation results for $n = 30$ subjects, based on 500 replications made with the same values of the covariates.

All the calculations were programmed in the R language, version 2.10.1. The function *MatrixExp* adopted in the program is from the R package *msm* contributed by Christopher Jackson. Convergence was considered to be obtained when the differences between the correct and previous parameter vector estimates were all less than 0.00001.

In Table 1, the *average estimate* is the average of the estimated parameters, over all simulations. Notice that these average estimated values are very close to the true values of the parameters. The *standard error* is the square root of the sample variance of the estimates and the *root MSE* is the square root of the mean squared error of the estimates. Notice that the values corresponding to β_1, β_3, λ_1, λ_3 and λ_5 are much larger than the standard error of the other estimates. The *coverage frequency* reports the percentage of times that the true parameter was contained in the corresponding 95% confidence intervals during all simulations. Notice that the coverage frequencies in Table 1 verifies the asymptotic normality property of the estimates.

The same experiment was repeated with $n = 100$ subjects, which was viewed as the considerably large sample situation. As expected, except for the average estimated values close to the true values and the coverage frequencies close to 95%, the smaller standard error and the root of mean-squared error are obtained for all estimates, for example, the standard errors of the estimates for mean are 0.0738, 0.0172, 0.0717 and 0.0174 respectively. This verifies the strong consistency of the estimates.

Table 1. Simulation Study Result Based on 500 simulations, Sample size 30, Repeated measurements 11

Parameter	True value	Average estimate	Standard error	Root MSE	Coverage frequency
β_1	5.0000	5.0043	0.1366	0.1366	95.6%
β_2	−2.0000	−2.0007	0.0348	0.0348	94.6%
β_3	4.0000	3.9816	0.1329	0.1341	94.0%
β_4	−2.0000	−2.0023	0.0332	0.0333	94.6%
γ_1	0.4000	0.3891	0.0383	0.0398	94.8%
γ_2	−0.0360	−0.0335	0.0098	0.0101	94.4%
γ_3	0.4000	0.3878	0.0396	0.0414	94.4%
γ_4	−0.0360	−0.0329	0.0102	0.0107	92.8%
γ_5	0.2000	0.2030	0.0377	0.0379	94.2%
γ_6	−0.0200	−0.0199	0.0095	0.0095	95.2%
γ_7	0.2000	0.2022	0.0387	0.0388	95.6%
γ_8	−0.0200	−0.0198	0.0099	0.0099	95.4%
λ_1	0.8000	0.7908	0.1595	0.1598	95.0%
λ_2	−0.0640	−0.0673	0.0246	0.0248	94.6%
λ_3	0.8000	0.8018	0.1631	0.1631	95.8%
λ_4	−0.0640	−0.0682	0.0234	0.0238	94.2%
λ_5	0.6000	0.6083	0.1173	0.1175	95.4%
λ_6	−0.0600	−0.0610	0.0184	0.0184	94.2%

References

Chiu, T.Y.M., Leonard, T., and Tsui, K-W. (1996). The Matrix-Logarithmic Covariance Model. *Journal of the American Statistical Association.* **91**, 198-210.

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

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika.* **86**, 677-690.

Comparisons of methods for dealing with missing data in longitudinal studies

Paola Zaninotto[1], Amanda Sacker[2]

[1] Department of Epidemiology and Public Health,1-19 Torrington Place, UCL, WC1E 7HB London, UK. email: p.zaninotto@ucl.ac.uk
[2] Institute for Social and Economic Research (ISER), University of Essex, UK

Abstract: In this study we focus on the analysis of data from the English Longitudinal Study of Ageing (ELSA), with incomplete time-dependent and time-independent variables. We compare three methods for dealing with missing data, such as Full Information Maximum Likelihood, Multiple Imputation under the normal model and MI using the two-fold Fully Conditional Specification.

Keywords: Missing data; Multiple imputations; Longitudinal data.

1 Aim

In this study we compare three methods for dealing with missing data that can be applied to longitudinal data, these are Full Information Maximum Likelihood (FIML), Multiple Imputation (MI) under the normal model (MVNI) and MI using the two-fold Fully Conditional Specification (FCS).

2 Methods

The FIML estimator estimates model parameters and standard errors using all available raw data, it does not impute or fill in missing values (Enders, 2001). Multivariate normal imputation (MVNI) assumes a joint multivariate normal distribution for all variables. For data sets with arbitrary missing patterns, a Markov Chain Monte Carlo (MCMC) method (Schafer 1997) can be used. A regression model is fitted for each variable with missing values, with the previous variables as covariates. Based on the fitted regression coefficients, a new regression model is simulated from the posterior predictive distribution of the parameters and is used to impute the missing values for each variable (Rubin 1987). The process is repeated sequentially for variables with missing values. The FCS approach differs from MVNI in that it does not start with the construction of a well-defined joint distribution for the variables to be imputed. FCS starts with a collection of univariate conditional distribution for each variable with missing data

in terms of all other variables. The main idea is that a univariate conditional model is constructed for each potentially missing variable which is appropriate to the type (i.e. logistic regression for binary variables). The other potentially missing variables are used as explanatory variables in each univariate imputation model. The conditional density for the jth missing variable (of p) and q repeated waves would be

$$f(X_{ij}|X_{i1},...,X_{i(j-1)},X_{i(j+1)},...,X_{ip}) \quad j=1...p \quad i=1...q \qquad (1)$$

Nevalainen et al (2009), proposed using an imputation strategy which is doubly iterative. At time i, X_i is imputed conditional on the same variable observed at time $i-1$ and $i+1$, and the other variables at time i. Univariate posterior draws are made one variable at a time by cycling through all p models given current values of the other variables. After sufficient cycles (10-20), the imputations are taken from one final cycle through the univariate model. One iteration runs over the variables $j=1....p$, called within time iterations. The past and future observations (X_{i-1} and X_{i+1}) are not imputed at this stage, they serve only the role of predictors in the imputation model. There is also a second imputation iteration over waves ($i=1...q$). The model is described in full details in Nevalainen et al (2009).

2.1 Description of the dataset

The data comes from the first 3 waves of the English Longitudinal Study of Ageing (ELSA), a panel study where 11,392 individuals aged 50 and over are followed and re-interviewed every two years (Marmot, 2003). We are interested in exploring the impact of Coronary Heart Disease (CHD) on Quality of Life (QoL) and to seek for possible gender differences. The data consist of one incomplete dependent variable (QoL) three complete covariates (Age, Sex, CHD) and six incomplete covariates (marital status, wealth, depression, physical activity, smoking status and alcohol consumption). For the purpose of this study we selected those people that at wave1 (2002-03) reported that in the two years preceding the interview were diagnosed with CHD and we compared them with a reference group of healthy individuals. The sample size is 4,496 in wave1; 3,465 in wave2 (2004-05) and 3,031 in wave3 (2006-07). Only 1,998 participants had complete data on all variables at the 3 waves (44.4% of the sample in wave1). Based on the resulting dataset of 1,998 individuals, we estimated the regression coefficients for QoL using the following random intercept model:

$$y_{ij} = \beta_{0ij} + \sum_{p=1}^{3} \beta_p x_{pj} + \sum_{p=1}^{8} \beta_p x_{pij} + u_j + e_{ij} \qquad (2)$$

where $e_{ij} \sim N(0,\sigma_e^2)$ and $u_j \sim N(0,\sigma_u^2)$.
Where x_j are time-invariant factors such as sex, CHD at wave1, and the

interaction term between CHD and sex; x_{ij} are time-varying factors such as age (a linear and quadratic term) marital status, depression, wealth, smoking status, alcohol consumption and physical activity. The targets parameters of interest we wish to cover are $\beta_1 = -1.47$ $\beta_2 = 1.10$ and $\beta 3 = -0.40$ (x_1= CHD, x_2= Sex and x_3=CHD*Sex).

2.2 Simulation strategy

To evaluate the performance of the three methods for missing data, an artificial simulation was set up, which was based on the real data for 1,998 individuals, the reference population for a complete data analysis. Simulations were done using 1000 replications. In every replication, 55.6% missing data was generated using random uniform numbers from the real dataset in order to reproduce the same probabilities of missingness as in the original study (of 4,496 individuals). If the rank of the random number was equal to or less than the percentage specified, the corresponding data point was deleted. Deletion was performed as follows: item non-response at wave1, drop-out, unit non-response and item non-response at wave2, drop-out and item non-response at wave3. Then each of the 1000 replications were analyzed as follow:

1. In Mplus to perform the FIML estimation
2. In SAS, using the MI procedure to generate 5 imputed datasets under the normality assumption
3. In Stata, to perform the two-fold FCS, generating 5 imputed datasets

For steps 2) and 3) the random intercepts regression coefficients and standard errors were estimated for each filled-in dataset, and the estimates were combined according to Rubin's rule (Rubin, 1987).

2.3 Evaluation criteria

The evaluation criteria used are bias of parameter estimates; standard deviation of estimates; Mean Square Error (MSE) of estimates; 95% confidence interval covarage, i. e. the proportion of the replications where a 95% CI covers the true parameter value; prevalence of the significance of coefficent at 5%.

3 Results and Discussion

Table 1 reports the results from the three methods. All methods show small bias effects and that the standard error of the parameter estimates are very close to target values. The three methods provide confidence intervals which successfully maintain their nominal 95% coverage. Results suggest that the

TABLE 1. Results

Population value	Coef.	SE	Bias	SD	MSE	95% CI Coverage	% Sig. coef.
CHD	-1.47	0.44					
Sex	1.10	0.27					
CHD*Sex	-0.40	0.68					
FIML							
CHD	-1.53	0.48	-0.06	0.18	0.04	1.00	1.00
Sex	1.17	0.29	0.06	0.11	0.02	1.00	1.00
CHD*Sex	-0.57	0.72	-0.17	0.28	0.11	1.00	0.00
2-Fold FCS							
CHD	-1.45	0.49	0.02	0.23	0.05	1.00	0.97
Sex	1.16	0.30	0.06	0.14	0.02	1.00	1.00
CHD*Sex	-0.42	0.74	-0.02	0.33	0.11	1.00	0.31
MVNI							
CHD	-1.42	0.48	0.05	0.20	0.04	1.00	0.99
Sex	1.02	0.29	-0.08	0.14	0.02	1.00	1.00
CHD*Sex	-0.43	0.72	-0.03	0.31	0.10	1.00	0.32

two-fold FCS estimates recover the population coefficients to an impressive extent.Nevertheless, given that the two-fold FCS is computationally intense and given the small gain over the estimates obtained using the FIML, researchers may prefer the FIML method.

References

Enders, C.K. (2001). The Performance of the Full Information Maximum Likelihood Estimator in Multiple Regression Models with Missing Data. *Educational and Psychological Measurement*, **61**,713-40.

Marmot, M.G (2003). *Health, wealth and lifestyles of the older population in England: the 2002 English Longitudinal Study of Ageing*. London: Institute for Fiscal Studies.

Nevalainen, J., Kenward M.G., and Virtanen, S.V. (2009). Missing Values in Longitudinal Dietary Data: A Multiple Imputation Approach Based on a Fully Conditional Specification. *Statistics in Medicine* **28**,3657-3669.

Rubin, D.B. (1987). *Multiple Imputations for Nonresponse in Surveys*. New York: Wiley.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data.*. London: Chapman & Hall.

Index

IWSM 2010 Sponsors

We are very grateful to the following organisations for sponsoring IWSM 2010.

- Biometrics and Data Management at Boehringer Ingelheim UK
- Chapman & Hall
- The Faculty of Information and Mathematical Sciences at the University of Glasgow
- Glasgow City Council
- Minitab
- Oxford University Press
- RSS
- SAS
- Statistics and Chemometrics at Shell
- Taylor & Francis
- Tunnocks
- Walkers