

**Proceedings of the
24th International
Workshop
on Statistical Modelling**

**July 20-24, 2009
Ithaca, NY**

**James G. Booth
(editor)**

Proceedings of the 24th International Workshop on Statistical Modelling.
Ithaca, NY, USA July 20-24, 2009
James G. Booth, editor
Ithaca, NY 2009.

Editor:

James G. Booth
Department of Biological Statistics and Computational Biology
Cornell University
1178 Comstock Hall
Ithaca, NY 14850
U.S.A.
jim.booth@cornell.edu

Scientific Programme Committee

- James Booth (Cornell University, USA)
- David Firth (Warwick University, England)
- Herwig Friedl (Graz University of Technology, Austria)
- Sonja Greven (Johns-Hopkins University, USA)
- Giles Hooker (Cornell University, USA)
- Clare Ferguson (University of Glasgow, Scotland)
- Joe Lang (University of Iowa, USA)
- Stefan Lang (Leopold-Franzens-University, Austria)
- M. Dolores Ugarte (Public University of Navarra, Spain)
- Matt Wand (University of Wollongong, Australia)
- Alan Watkins (University of Swansea, Wales)
- Martin Wells (Cornell University, USA)



Preface

This proceedings volume consists of papers submitted to the 24th annual International Workshop on Statistical Modeling (IWSM) which is being hosted by the Department of Statistical Sciences at Cornell University.

The workshop originated out of two GLIM conferences in the U.K., in London (1982) and Lancaster (1985), which attracted many statisticians interested in generalized linear modeling. The inaugural workshop was held in 1986 in Innsbruck, Austria, and subsequent workshops were held in Italy, France, The Netherlands, Germany, Belgium, England, Switzerland, USA, Spain, Denmark, Greece, Australia and Ireland. This year's workshop will be attended by participants from 20 countries from six continents. The 2010 IWSM will be held in Glasgow, Scotland.

The workshop has evolved over the past two decades into a broad-based conference focused on applied statistical modeling motivated by real-life data. We thank all the authors who have contributed to this proceedings volume which includes an impressive collection of work on novel statistical methodology developed for a diverse set of problems and application areas. We also thank the Scientific Program Committee, consisting of James Booth, David Firth, Herwig Friedl, Sonja Greven, Giles Hooker, Claire Ferguson, Joe Lang, Stefan Lang, Lola Ugarte, Matt Wand, Alan Watkins and Marty Wells, who reviewed all the submitted abstracts. This year the workshop will include a poster session and approximately 40 contributed talks given in a plenary session. In addition, five special invited talks will be given by Brian Caffo (Johns-Hopkins), Rebecca Doerge (Purdue), Goran Kauermann (Bielefeld), Emmanuel Lasaffre (Erasmus University), and David Ruppert (Cornell). We are also delighted that Murray Aitkin agreed to give a short course on "Bayesian Model Selection" prior to workshop.

The workshop provides a forum for researchers from different countries to exchange ideas and foster collaborations on wide-ranging subjects for which statistical modeling plays a key role. The journal "Statistical Modeling" was started in 2000, and in 2003 the Statistical Modeling Society (www.statmod.org) was founded as an umbrella organization for the workshop, the journal and related activities.

Financial support for the Ithaca workshop came from several sources including a grant from the National Security Agency. We also acknowledge generous contributions from the Department of Statistical Sciences, the College of Industrial and Labor Relations and the College of Agriculture and Life Sciences at Cornell.

We are extremely pleased to be hosting the IWSM here in Ithaca. We sincerely hope that the participants have a rewarding experience and enjoy their visit to the Finger Lakes region of New York state.

Jim Booth
July 2009



Part 1. Invited papers

B. CAFFO ET AL. A survey of statistical methods for non-invasive measurement of connectivity in the human brain	3
R. W. DOERGE AND K. KIM Statistical Challenges in Modeling Expression Quantitative Traits	13
G. KAUEMANN Penalized Spline Estimation and Mixed Models – A Flourishing Statistical Partnership	23
E. LESAFFRE AND D. DECLERCK Statistical Modeling in Oral Health Research	39

Part 2. Contributed papers

M.A. AMARAL-TURKMAN AND K.F. TURKMAN Hierarchical space-time model for fire ignition and percentage of land burned by wildfires	49
H. BAR AND E. SCHIFANO Bayesian Approaches for Random Effects Models in Microarray Analysis	53
B. BLAS ACHIC, H. BOLFARINE, AND H. LACHOS Diagnostic on controlled calibration model with replicates on the both variables	61
A.W. BOWMAN ET AL. Mapping brain activity through spatiotemporal smoothing	69
R. CABALLERO, A. HERMOSO-CARAZO, AND J. LINARES-PÉREZ Least-squares quadratic estimation in uncertain observation systems with different uncertainty probabilities	75
C. CAMARDA, P. EILERS, AND J. GAMPE Modelling trends in digit preference patterns	81
D. CHAKRABARTY CHASSIS - Nonparametric Bayesian Estimates of Gravitational Potential and Phase Space Density Function	89
J. CIERA, B. SCARPA, AND D. DUNSON Fast Bayesian Functional Data Analysis: Application to basal body temperature data	96
R. COLOMBI AND S. GIORDANO Multi Edge Graphs for Multivariate Markov Chains	102
D. COSTAIN Matched case-control data: a Bayesian partition modelling approach to mapping residual spatial variation in disease risk	110
C. CUNNINGHAM AND J. BOOTH A Bayesian Approach to Analysis of Covariance for Split-Plot Designs	118
T. ECONOMOU ET AL. A latent structure model for high river flows	125
P. EILERS Deconvolution of Spike Trains Using an L_0 Penalty	130
L. FABIO, G. PAULA, AND M. DE CASTRO A generalized random intercept log-gamma-Poisson model	138
S. GREVEN AND T. KNEIB Marginal and Conditional Akaike Information Criteria in Linear Mixed Models	142
L. HAINES AND K. LEASK Exponentially Weighted Poisson Models	150

B. HANLON ET AL. A Bayesian Approach for Evaluating Drug Efficacy using Fecal Egg Count Data	154
G. HELLER ET AL. Randomly stopped sum models: a hydrological application	160
M. HODGE, J. BROWN, AND M. REALE Improving the Calculation of Fix-Rate Bias in Automated Telemetry Systems	168
S. HUZURBAZAR AND J. BARBER Bayesian Modelling of Grainsize Distributions	174
N.K. JAJO AND K.M. MATAWIE Eigenvalues Application in Robust outlier Detection	182
A. KOMÁREK ET AL. Prediction of binary response using multivariate longitudinal profiles: Study on chronic hepatitis B patients	187
I. KOSMIDIS The iterative adjustment of the responses for the reduction of bias in binary regression models	193
D. LEE AND M. BURBÁN Nested B -spline bases: an efficient method for spatio-temporal smoothing	202
D. LIN ET AL. Nonparametric Detection of Outliers in Multivariate Data Streams	209
G. MACKENZIE AND J. XU Space -Time Clustering Revisited	217
G. MARRA AND S. WOOD Confidence Intervals for Generalized Additive Model Components	223
N. MARTÍN AND L. PARDO Fitting DNA sequences through loglinear modeling with linear constraints	231
E. MARTÍNEZ-GÓMEZ AND G. BABU A statistical model for the relation between exoplanets and their host stars	237
T. MARY-HUARD, É. LEBARBIER AND S. ROBIN A clustering method for ordered variables to detect up-correlated genomic regions	243
E. MOLANES-LÓPEZ AND E. LETÓN Empirical likelihood based approach for the inference of the Youden index and associated threshold	247
G. NEUBAUER AND G. DJURAŠ A Beta-Poisson Model for Underreporting	255
G. NEUBAUER, M. DVORZAK, AND H. WAGNER Bayesian Estimation of a Beta-Poisson Model	261
A. NOUFAILY AND C. JONES On a Family of Distributions in the Context of Quantile Regression	267
G. OSKROCHI ET AL. An Application of the Multivariate Linear Mixed Model to the Analysis of Shoulder Complexity: EMG Measurements in Breast Cancer Patients	273
H. PARK AND S. HONG A test procedure for right censored data under the additive model	281
C. PRAMANIK Computation of Agriculture	289
P. PUIG ET AL. Statistical models for retinal image matching	295

R. RADICE AND G. MARRA Instrumental Variable Estimation for Generalized Additive Models	300
R. RIPPE ET AL. Improved SNP genotyping using model-based calibration	306
O. ROSEN, S. WOOD AND R. KOHN Modeling Time Varying Parameter Models Using Mixtures	314
C. RUSSO, R. AOKI, AND G. PAULA Assessment of variance components in elliptical nonlinear models for correlated data	322
S. SCHNABEL AND P. EILERS Non-crossing smooth expectile curves	330
K. SELLERS AND G. SHMUELI A Regression Model for Count Data with Observation-Level Dispersion	337
E. SILVA, V. GUERRERO AND D. PEÑA Smoothing two-dimensional mortality rates with a given percentage of smoothness	345
A. SIMPKIN ET AL. An Additive Penalty Approach to Derivative Estimation of Noisy Data	351
A. VAN DEN HOUT AND F. MATTHEWS Investigating smoking behaviour as a risk factor for stroke-free life expectancy	359
C. VILLEGAS, G. PAULA, AND V. LEIVA Birnbaum-Saunders random intercept models with censored data	365
J. XU AND G. MACKENZIE Modelling of covariance structure in constrained marginal models for longitudinal data	369



Part 1
Invited papers



A survey of statistical methods for non-invasive measurement of connectivity in the human brain

Brian Caffo¹, Haley Hedlin¹, Suresh Joel¹, Stewart Mofstovsky¹, Jim Pekar¹, and Susan Spear-Bassett¹

¹ Johns Hopkins Bloomberg School of Public Health

Abstract: New developments in cognitive neuroscience focus on the human brain's functional integration. This is the study of the functional interactions, causes and the anatomical hierarchical circuitry of the brain, the variability in these quantities in the population and their impact on brain function and health. Functional connectivity in particular is defined as correlations between spatially remote neurological events. In this manuscript, we give a brief overview of singular value decomposition methods for statistically estimating brain networks associated with functional connectivity. We give a brief review of connectivity and eigen-value decompositions for estimating functional connectivity using functional magnetic resonance imaging.

1 Introduction

The classical study of brain function focused on the idea of functional specialization (Friston et al., 2007; Friston, 2005; Horwitz, 2003). That is, associating various cognitive and motor functions to specific areas of the brain and the degree and variation of specialization across populations. The history of such exploration predates modern imaging by investigating the behavior of stroke patients with localized lesions or persons with other localized brain injuries. A relatively more recent topic is the idea of functional integration. This is the study of interactions and how various areas of the brain cooperate, how these interactions are mediated, the variability of these interactions in the population and their association with brain function and disease. This area is referred to as the study of brain connectivity. In this manuscript we overview connectivity and some statistical methods used to analyze connectivity data from functional magnetic resonance imaging (fMRI).

2 Brain connectivity

Being under rapid current development and attempting to describe a complex idea, brain connectivity remains a difficult concept to formally define (Horwitz, 2003). Current nomenclature divides connectivity into three overlapping, biologically related areas: anatomical, functional and effective (Friston et al., 2007;

Friston, 2005; Horwitz, 2003; Stephan et al., 2008; Friston, 1994). **Anatomical connectivity** refers to actual axonal, synaptic and cortical connections in the brain. Conceptually, anatomical connectivity refers to the underlying circuitry of the brain. **Functional connectivity** is a concept that is typically defined statistically as *correlations between spatially remote neurological events* (Friston, 1994). We broaden this definition to consider functional connectivity as potentially lagged temporal synchronicity in measured brain function. **Effective connectivity** is causal in nature, representing the impact of neuronal systems on each other.

While it is tempting to consider functional connectivity as being a consequence of effective connectivity and both being a consequence of anatomical connectivity, the ideas are more complex, especially considering the level of detail attainable for each of the connectivity concepts with current measurement techniques. Instead, the three areas of connectivity must be thought of as related, but distinct, concepts and their quantification inherently tied to the measurement technique used for estimation (Horwitz, 2003).

Another difficulty in this area lies in creating statistical methods to analyze connectivity data. Many of the technologies create data sets containing millions of measurements. Moreover, effective and functional connectivity inherently require higher order summaries than means or main effects, such as correlations and interactions. Thus if N measurements are recorded, often N^2 quantities are under consideration for connectivity studies. Below we discuss eigen-value decompositions for analyzing *functional connectivity* for fMRI.

3 fMRI

Functional MRI has been nothing short of revolution in cognitive neuroscience. Functional MRI data is a time-series of MRI images, most commonly targeting the BOLD (blood oxygenation level dependent) signal. The central dogma of fMRI is the increase in regional cerebral blood flow from localized neuronal activity. Thus in areas of active use, the increase in blood flow results in an increase in diamagnetic oxyhemoglobin over paramagnetic deoxyhemoglobin. This effect is the biological basis for the signal in BOLD fMRI.

Much of the early seminal work in fMRI focused on the study of activation maps associated with a blocked experimental paradigm, working off of the idea of functional specialization. Typically, these maps were thresholded and compared to the expected Euler characteristics from Gaussian random fields (GRF) (Worsley, 1994; Friston et al., 1993; Worsley et al., 1996; Friston et al., 1995; Worsley, 1996). This work has been extended to event related paradigms and forms the core of the popular statistical package SPM Friston et al. (2007). Data resampling methods are also popular and avoid some of the distributional concerns of GRF methods (see Hayasaka and Nichols, 2003; Nichols and Holmes, 2002).

In addition to the study of paradigm-related activity, the hypothesis of a default mode brain network and the idea of studying functional connectivity using

fMRI have begun to make important inroads in cognitive neuroscience (Biswal et al., 1995; Greicius, 2003, 2004). In these studies, a subject is asked to sit at rest in the fMRI scanner with their eyes closed. While controversy exists over the potential role of confounding variables in the existence and importance of default-mode brain networks, we note that they have shown inter-subject differences between diseased and non-diseased subjects (Greicius, 2004).

4 Eigenimages

Let \mathbf{X} be a T by V matrix with each column representing a normalized and smoothed fMRI image of non-background voxels, where T is the number of time points and V is the number of non-background voxels. Typically, V on the order of tens or hundreds of thousands is much larger than T , which is typically on the order of hundreds. Let $\tilde{\mathbf{X}}$ be the centered version of \mathbf{X} over time and voxels; that is, $\tilde{\mathbf{X}} = \{\mathbf{I} - \mathbf{1}(\mathbf{1}^t\mathbf{1})^{-1}\mathbf{1}^t\}\mathbf{X}\{\mathbf{I} - \mathbf{1}(\mathbf{1}^t\mathbf{1})^{-1}\mathbf{1}^t\}$ for identity matrix \mathbf{I} and vector of ones $\mathbf{1}$. Left multiplication by $\{\mathbf{I} - \mathbf{1}(\mathbf{1}^t\mathbf{1})^{-1}\mathbf{1}^t\}$ takes voxel-wise residuals after regressing out an intercept-only model. This subtracts off the voxel-specific mean across time. Right multiplication subtracts the average across voxels for each time point. That is, this regression subtracts the time series obtain by averaging across all (non-background) voxels from every voxel-level time series. Such “global-signal regression” ideally removes physiological effects and other temporal nuisance variables that are not localized. In the analysis of functional connectivity, the left multiplication is always performed while there remains some discussion on the utility of global-signal regression. The **global connectivity matrix** is defined as $\mathbf{C} = \tilde{\mathbf{X}}^t\tilde{\mathbf{X}}/T$. That is element (v, v') of \mathbf{C} is the correlation between voxels v and v' over time. This matrix completely summarizes synchronous functional connectivity information in the image. However, it is not parsimonious, containing V choose 2 unique non-diagonal elements. Moreover, the rank of the matrix is no larger than T , hence it contains many redundant elements. Finally, individual voxel-level connectivity is not of interest. Instead, large patterns of synchronous behavior are. Therefore, parsimonious methods to summarize this $V \times V$ matrix are needed. Singular value decomposition (SVD) analysis of $\tilde{\mathbf{X}}$ helps study principal directions of variations and their geometric groupings in the data. Moreover, it is a useful first dimension reduction step for later analysis. Let

$$\tilde{\mathbf{X}} = \mathbf{W}\mathbf{D}\mathbf{U}^t, \quad (1)$$

for $T \times T$ matrix \mathbf{W} , diagonal matrix of eigen-values \mathbf{D} and $V \times T$ matrix \mathbf{U} so that $\mathbf{U}^t\mathbf{U} = \mathbf{W}^t\mathbf{W} = \mathbf{I}$. Dimension reduction occurs by retaining only a few of the columns of \mathbf{U} and rows of \mathbf{W} corresponding to the largest eigen-values. With this decomposition, the columns of \mathbf{W} are time-series often referred to as eigenvariates while the columns of \mathbf{U} are images referred to as eigenimages (Friston et al., 2007). The SVD can be either thought of as representing every voxel's time series as a mixture of the orthogonal time series from \mathbf{W} or

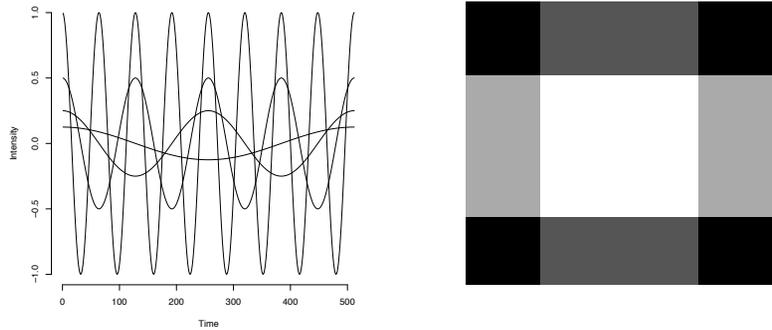


FIGURE 1. Example eigenvariates (left panel) and eigenimages (right panel) used for simulation.

every image at each time as a mixture of eigenimages from \mathbf{U} . Unlike other methods, such as ICA, the SVD does not distinguish between which of these representations is desired.

Another interpretation of the SVD is to relate it to voxel-wise regression analysis. Specifically, $\tilde{\mathbf{X}}$ is the residual fMRI data having regressed out a voxel-wise intercept. Then consider the regression model $\tilde{\mathbf{X}} = \mathbf{H}\beta + \epsilon$ where \mathbf{H} is a $T \times P$ design matrix with columns comprised of orthonormal time series and β is a $P \times V$ matrix of voxel-wise coefficients and ϵ is an error term. Then consider both \mathbf{H} and β being estimated via least squares, unlike typical voxel-wise regression models where only β is estimated and \mathbf{H} is specified. Here the problem does not have a unique solution, since given an estimate of \mathbf{H} any rotation yields an equivalent model and fit. However, if one is to add the constraint that \mathbf{H} is estimated so that its first column, say \mathbf{h}_1 , is the unit vector so that $\mathbf{h}_1^t \tilde{\mathbf{X}}$ is most variable, the second one the unit vector orthogonal to \mathbf{h}_1 being second most variable and so on, one obtains that the estimate of \mathbf{H} is the first P columns of \mathbf{W} and the estimate of β is the first P rows of $\mathbf{D}\mathbf{U}^t$.

The SVD implies the eigen-value decomposition on the global connectivity matrix given by $\mathbf{C} = \mathbf{U}\mathbf{D}^2\mathbf{U}^t/T$. Thus, the global connectivity matrix can be written as $\mathbf{C} = \sum_j D_j^2 \mathbf{u}_j \mathbf{u}_j^t / T$ where \mathbf{u}_j is column (eigenimage) j from \mathbf{U} and D_j is the j^{th} eigen-value; that is, the connectivity matrix can be written as a sum of outer-products of the eigenimages. The eigenimages form a basis of which every image in the time series is a linear combination. These images are often thought of as distributed brain networks (Friston et al., 2007). Being a decomposition of the global connectivity matrix, these networks are informative regarding functional brain connectivity. One typically plots the eigenimages overlaid on anatomical templates to assess connectivity.

To illustrate we created a simple simulation setting. Figure 1 shows four functions (cosines of different periods left panel) and an image space (right panel)

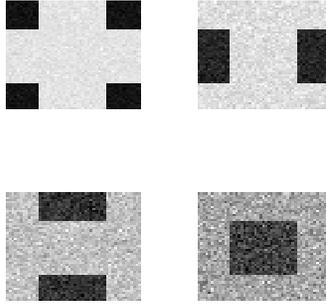


FIGURE 2. Eigenimages from simulated data.

of four networks (one for each shade of gray). There are $50 \times 50 = 2,500$ pixels and 512 time points. The first function is added to mean zero Gaussian noise with a standard deviation of .5 in the darkest region, the second is added to white noise in the next darkest, and so on. The data matrix was then $512 \times 2,500$ and was centered on both rows and columns. Then the SVD was applied to the centered data matrix. The scree plot (not shown) demonstrated that the first four eigen-values cover 40% of the variation in the data. The first four eigenvariates match the true functions nearly exactly. The first four eigenimages correspond well with the four brain networks (see Figure 2).

In terms of calculation, note the eigen-value decomposition of $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^t/V = \mathbf{W}\mathbf{D}^2\mathbf{W}^t/V$. Here, provided global-signal regression has been performed, $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^t$ is the $T \times T$ matrix representing correlation across voxels at each time point. Having only a few thousand entries, this low-dimensional eigen-value decomposition is easily calculated using standard techniques. Then, one can calculate $\mathbf{U}^t = \mathbf{D}^{-1}\mathbf{W}^t\tilde{\mathbf{X}}$. This way, the high dimensional $V \times V$ global connectivity matrix never needs to be calculated to evaluate the brain networks supplied by the SVD.

In our example we show the eigenvariates and images for a subject performing a Stroop task (Banich et al., 2000; Redgrave et al., 2008). Figure 3 shows the first nine eigenvariates for this fMRI time series plotted by the time of repetition. The first clearly represents an across the board mean shift in the fMRI signal over time. The second represents a linear decrease or increase in signal intensity. The fourth and fifth components are periodic. Figure 4 displays an example eigenimage overlaid on a template image.

5 Limitations of the SVD

As an approach for creating brain networks the SVD has many limitations. First, it only considers linear correlations. Moreover, it can only consider syn-

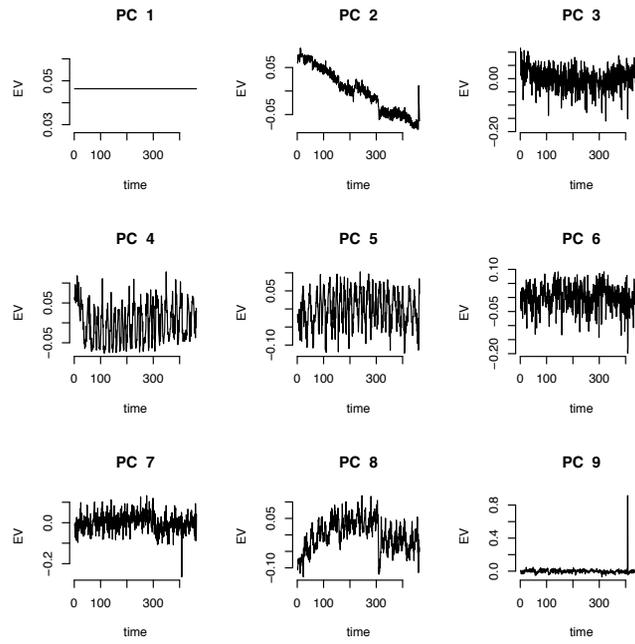


FIGURE 3. Eigenvariate time series for fMRI Stroop task image.

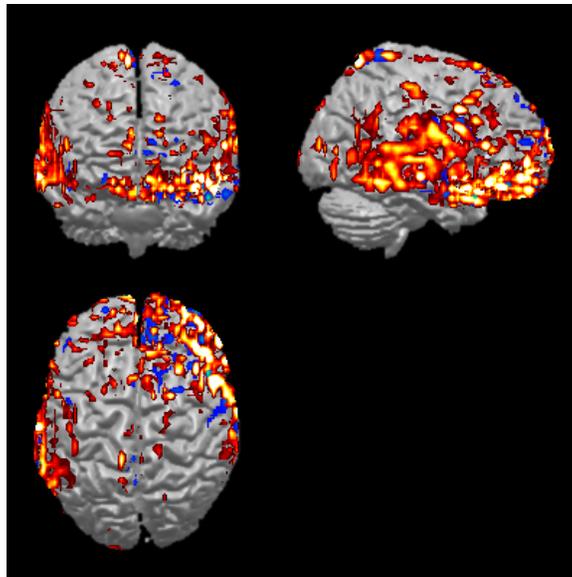


FIGURE 4. Example eigenimages from the Stroop data example.

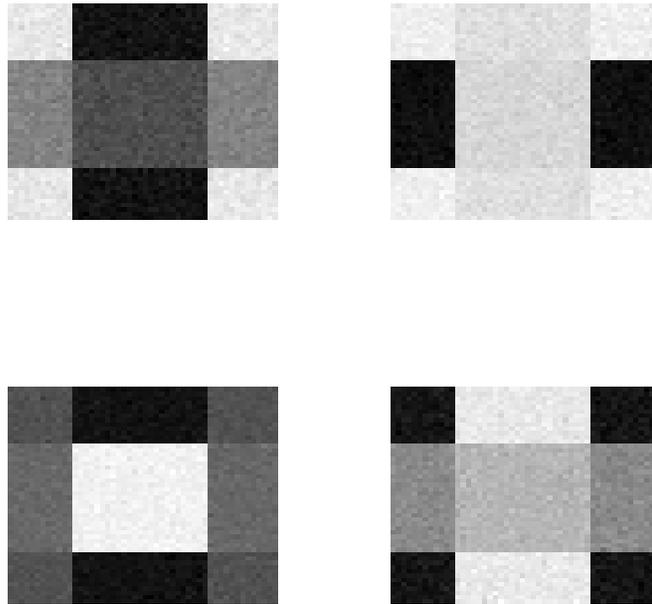


FIGURE 5. Incorrect networks obtained from SVD.

chronous signals, hence presuming consistent haemodynamics throughout the brain, which is false for reasons of both biology and instrumentation (see Lindquist and Wager, 2007). Moreover, the SVD is only a rotation of the data and choosing the decomposition to maximize variability can miss important features represented by higher order moments. As an example, our simple simulation can be used to highlight the limitations of SVD analysis of fMRI data. Simply by allowing the four cosine functions to have equal amplitude (they are decreasing in above example), the first four eigenimages are as in Figure 5. There, because of the lack of a decrease in variation of the underlying signals, the brain networks are mixed together.

Independent components analysis offers a solutions to some of the limitations of the SVD. While the SVD focuses on data-level orthogonality, ICA is a general factor analysis model that instead considers model-based independence and non-normality (Hyvärinen et al., 2001). ICA has been used extensively in the analysis of fMRI data (see Calhoun et al., 2003; McKeown et al., 1998).

We close by noting that this manuscript has only touched on the tip of the iceberg for the study of functional connectivity. We anticipate that the study

brain connectivity in general will remain one of the most exciting new fields of scientific inquiry in the near future. One of the most exciting areas is the integration of different connectivity measures in combined analyses. For example, fMRI and DTI are being combined in unified approaches (Skudlarski et al., 2008; Pearlson and Calhoun, 2007).

Bibliography

- M.T. Banich, M.P. Milham, R. Atchley, N.J. Cohen, A. Webb, T. Wszalek, A.F. Kramer, Z.P. Liang, A. Wright, J. Shenker, et al. fMRI studies of stroop tasks reveal unique roles of anterior and posterior brain systems in attentional selection. *Journal of Cognitive Neuroscience*, 12(6):988–1000, 2000.
- B. Biswal, F.Z. Yetkin, V.M. Haughton, and J.S. Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic Resonance Medicine*, 34(4):537–41, 1995.
- V.D. Calhoun, T. Adali, L.K. Hansen, J. Larsen, and J.J. Pekar. ICA of functional MRI data: an overview. In *Fourth International Symposium on Independent Component Analysis and Blind Source Separation*, pages 281–288, 2003.
- K Friston, J Ashburner, K Stefan, T Nichols, and W Penny, editors. *Statistical Parametric Mapping The Analysis of Functional Brain Images*. Academic Press, 2007.
- K.J. Friston. Functional and effective connectivity in neuroimaging: a synthesis. *Human Brain Mapping*, 2, 1994.
- K.J. Friston. Models of brain function in neuroimaging. *Annual Reviews of Psychology*, 56:57–87, 2005.
- KJ Friston, KJ Worsley, RSJ Frackowiak, JC Mazziotta, and AC Evans. Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1(3):210–220, 1993.
- KJ Friston, AP Holmes, KJ Worsley, JB Poline, CD Frith, RSJ Frackowiak, et al. Statistical parametric maps in functional imaging: a general linear approach. *Hum Brain Mapp*, 2(4):189–210, 1995.
- M.D. Greicius. Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences*, 100(1):253–258, 2003.
- M.D. Greicius. Default-mode network activity distinguishes alzheimer’s disease from healthy aging: Evidence from functional MRI. *Proceedings of the National Academy of Sciences*, 101(13):4637–4642, 2004.

- S. Hayasaka and T.E. Nichols. Validating cluster size inference: random field and permutation methods. *Neuroimage*, 20(4):2343–2356, 2003.
- Barry Horwitz. The elusive concept of brain connectivity. *NeuroImage*, 19(2): 466–470, 2003. ISSN 1053-8119. doi: DOI: 10.1016/S1053-8119(03)00112-5. URL <http://www.sciencedirect.com/science/article/B6WNP-48J44GM-P/2/a63e7b52e8ff3c8972e71d87743c57df>.
- A. Hyvärinen, J. Karhunen, and E. Oja. Independent component analysis. *John and Wiley, New York*, 2001.
- M.A. Lindquist and T.D. Wager. Validity and power in hemodynamic response modeling: A comparison study and a new approach software available for free download from: <http://www.columbia.edu/cu/psychology/tor>. *Human brain mapping*, 28(8), 2007.
- M.J. McKeown, S. Makeig, G.G. Brown, T.P. Jung, S.S. Kindermann, A.J. Bell, and T.J. Sejnowski. Analysis of fMRI data by blind separation into independent spatial components. *Human Brain Mapping*, 6(3), 1998.
- T.E. Nichols and A.P. Holmes. Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, 15(1):1–25, 2002.
- G.D. Pearlson and V. Calhoun. Structural and functional magnetic resonance imaging in psychiatric disorders. *Canadian Journal of Psychiatry*, 52(3):158, 2007.
- G.W. Redgrave, A. Bakker, N.T. Bello, B.S. Caffo, J.W. Coughlin, A.S. Guarda, J.E. McEntee, J.J. Pekar, S.P. Reinblatt, G. Verduzco, et al. Differential brain activation in anorexia nervosa to fat and thin words during a stroop task. *NeuroReport*, 19(12):1181, 2008.
- P. Skudlarski, K. Jagannathan, V.D. Calhoun, M. Hampson, B.A. Skudlarska, and G. Pearlson. Measuring brain connectivity: diffusion tensor imaging validates resting state temporal correlations. *NeuroImage*, 43(3):554–561, 2008.
- K.E. Stephan, J.J. Riera, G. Deco, and B. Horwitz. The brain connectivity workshops: Moving the frontiers of computational systems neuroscience. *NeuroImage*, 42(1):1–9, 2008.
- K Worsley. Local maxima and the expected Euler characteristic of excursion sets of χ^2 , F, and t fields. *Advances in Applied Probability*, 26:13–42, 1994.
- K.J. Worsley. The geometry of random images. *Chance*, 9:27–40, 1996.
- KJ Worsley, S. Marrett, P. Neelin, AC Vandal, KJ Friston, and AC Evans. A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, 458:73, 1996.



Statistical Challenges in Modeling Expression Quantitative Traits

R.W. Doerge¹ and Kyunga Kim²

¹ Department of Statistics, Purdue University, West Lafayette, IN 47907 USA

² Department of Statistics, Seoul National University, Seoul 151-747, Korea

Abstract: Complex traits are typically controlled by multiple genes that are organized into networks that may behave differently under varying environmental conditions. Identifying the regions of the genome, or genes, associated with complex traits has been a long standing interest in the scientific community. Using microarray technology, complex traits can be molecularly dissected and the determinants of expression level polymorphism (ELP; the quantified variation in mRNA) established for the purpose of providing a systems biological way to investigate gene networks.

While progress has been made in both theory (quantitative trait locus, QTL, analysis) and technology (microarrays) limited progress has been made in unifying QTL mapping and microarray technology for the purpose of molecularly dissecting a complex trait. Toward this end, a framework for designing, understanding, and analyzing ELP experiments is proposed with an eventual focus on constructing gene regulatory networks. A novel multivariate mixture linear model (MMLM) with mixed effects is presented for the purpose of mapping genetic determinants of ELPs. Practical recommendations for future ELP studies are given, and suggestions regarding the incorporation of ELP mapping results to gene network construction are made.

Keywords: gene expression; quantitative trait locus (QTL) mapping; eQTL; genomics.

1 Introduction

The underlying mechanism of inheritance has been studied through various approaches in many areas of science. While each area in genome science (e.g., genetics, genomics, proteomics, etc.) provides information on components that regulate systems of genes (or gene networks), recent progress in technology and analytic methodology continues to provide ways to unravel complexity at a systems biological level (Jansen, 2003). There has been growing interest in incorporating the methodologies of genetics and genomics (Jansen and Nap, 2001; Doerge, 2002) to pursue a more comprehensive dissection of complex traits and their genomic architecture. An approach relying on both QTL methodology and gene expression data from

microarray experiments was outlined by Doerge (2002) who introduced the concept of expression level polymorphism (ELP; the quantitative variation in mRNA transcript measurements (Doerge, 2002; Kim et al., 2005).

We present an approach, called expression level polymorphism (ELP) analysis (Kim, 2007) that identifies and locates genetic determinants of the quantitative variation in mRNA transcripts (i.e., ELPs) by integrating quantitative trait locus (QTL) analysis and gene expression analysis in a statistically powerful manner. The combination of two methodologies gives rise to numerous technical and statistical issues, including experimental design, sample size and replication, normalization of gene expression data, maximum likelihood estimation in multivariate linear mixed-effects models, and corrections for multiple testing.

1.1 Microarray Technology

All microarray technologies are based on the concept of pairing events (or hybridization) between two single-stranded cDNA sequences, and share commonalities that include array fabrication, mRNA preparation and hybridization, and (visual and numeric) assessment of mRNA abundances. During the array fabrication process (biological or synthetic) genetic material (called probes) representing genes are affixed, in a grid-like pattern, to an array or a chip that consists of a solid surface (e.g., glass, silicon or plastic). The GeneChip technology that we rely on here is a product of Affymetrix (2002) and is an example of oligonucleotide microarrays that are commercially manufactured in large quantities. The probe sets are artificially produced, usually by in-situ synthesis based on photolithography (Lipshutz et al., 1999).

Experiments are conducted by extracting mRNA samples from an organism(s) at certain time points with and/or without a treatment of interest. For the Affymetrix GeneChip (Affymetrix, 2002) each sample is hybridized to a single GeneChip. Relying on complementation of base pairs, the expressed genes in each sample will find its complement, or partner, on the GeneChip and bind. Those mRNA that are not represented on the GeneChip, and have no partner, do not bind and are not recognized. After hybridization, images are collected by scanning the fluorescence levels of mRNA bound to each spot on the slides. The numeric mRNA abundances are gained from the raw images via image processing that includes gridding, segmentation, intensity acquisition, background correction, and quality control.

1.2 Quantitative Trait Locus Mapping

Most inherited traits (e.g., developmental, disease resistance, etc.) are controlled by many genes, and almost always interact with the environment

(Lynch and Walsh, 1998). These are called complex traits, or quantitative traits, and demonstrate continuous variation. By contrast, traits involving only a single gene are called monogenic, or Mendelian, traits. Understanding the genetic basis of complex traits is a major interest to scientists, and despite the rich statistical literature that has contributed to the analysis and dissection of complex traits, limited progress has been made in unraveling this mystery (Risch, 2000; Anholt and Mackay, 2004).

Quantitative trait locus (QTL) analysis is an approach that relies on standard statistical theory to understand the relationship between phenotypic (i.e., trait) variation and its genetic basis for the purpose of identifying chromosomal regions associated with traits. These chromosomal regions are typically large, may contain hundreds of genes, and are known as quantitative trait loci (QTL). QTL are typically located relative to a genetic map (chromosomally ordered genetic markers that act as fence posts in the genome), and explain only a small portion of the total (phenotypic) variation of the trait.

Since early QTL studies by Sax (1923) and Thoday (1961) genetic markers have been used to identify QTL based on a simple idea that the existence of a QTL close to a genetic marker would result in the association between the marker and the trait. The simplest way to detect QTL is single-marker analysis, which compares the phenotype (trait) distribution of the marker genotype classes at each marker locus, via t-test, regression, or the LOD (log odds ratio) score test (Doerge et al., 1997).

The availability of detailed genetic maps has enabled advanced QTL methodologies. Interval mapping (IM) is a popular maximum likelihood (ML) (Lander and Botstein, 1989) approach that uses consecutive intervals of a genetic map to locate single QTL associated with a single trait. The logarithm of the odds favoring linkage (LOD) or the likelihood ratio test (LRT) is evaluated at incremental positions across the genetic map and the results displayed as a profile map of test statistics that indicate the existence of a single potential QTL. When a LOD/LRT score is larger than a proper threshold, the corresponding QTL is identified with statistical significance. Significance thresholds for QTL mapping are typically determined using a permutation-based method (Churchill and Doerge, 1994).

When QTL are too close together, or interact (epistasis), the statistical power of single-QTL approaches (e.g., single-marker analysis and IM) is limited. As such, interval mapping was extended to composite interval mapping (CIM) (Zeng, 1993, 1994), multiple-QTL-model mapping (MQM) (Jansen, 1993, 1994), and multiple interval mapping (MIM) (Kao and Zeng, 2002; Kao et al., 1999) in attempts to boost the statistical power of single trait analyses. Because most QTL mapping studies measure numerous quantitative traits it is desirable to identify QTL associate with multi-

ple correlated traits. Multiple trait QTL methodologies (Jiang and Zeng, 1995) enable the study of the genotype-environment interaction via the statistical equivalency between the assessment of the same trait in multiple environments and the assessment of multiple correlated traits in the same environment (Falconer, 1952). Certainly, many other approaches have been developed since these extensions, but are too numerous to mention here.

QTL mapping relies on information from the phenotype and marker genotypes of each individual in a segregating population. By contrast, gene expression analysis uses the genome-wide patterns of mRNA transcription under various treatments to reveal the relationship between mRNA transcription and phenotype (or response to the treatment). Combining QTL analysis with microarray expression analysis enables the dissection of gene expression as a complex trait. We present a powerful approach to identifying and locating genetic determinants of the quantitative variation in mRNA transcript abundance (ELP) for the purpose of understanding gene regulatory networks and their relationship to complex traits. The combination of these two methodologies introduces numerous technical and statistical issues, including experimental design, sample size and replicates, and maximum likelihood estimation in multivariate linear mixed-effects models.

2 Expression Level Polymorphism Interval Mapping

Natural variation, including genetic variation of expressed genes when measured across individuals in an experimental population is referred to as expression level polymorphism or ELP (Doerge, 2002). By decomposing ELP into genetic and nongenetic factors, and their interactions it is possible to identify genomic regions affecting gene expression levels and to further characterize a complex trait via ELP for genes which are measured over various experimental conditions that presumably reflect a pathway involved in the realization of the trait.

ELP interval mapping (EIM) is adapted from traditional interval mapping (Lander and Botstein, 1989). EIM is based on a multivariate mixture linear model (MMLM) with mixed effects that represent all sources of ELPs, including nongenetic sources associated with using microarrays, such as array effect. In MMLM, a multivariate normal mixture is assumed for the distribution of expression levels of two structural genes under two treatment conditions. Maximum likelihood parameter estimates are computed through the expectation conditional maximization (ECM) algorithm (Meng and Rubin, 1993). Similar to interval mapping, EIM employs the likelihood ratio test (LRT) statistic to identify genetic determinants of ELP and to detect the gene-by-treatment interaction. Statistical significance is assessed via a permutation-based threshold is (Churchill and Doerge, 1994).

2.1 Statistical Model

Suppose that n individuals are sampled from a recombinant inbred line (RIL) population where each genetic marker can assume one of two states (or genotypes). Each individual is genotyped for a number of genetic markers, and m transcript measurements for a specific gene are gained from a microarray experiment. A genetic map is available, and the genotype of each marker is recorded as 0 and 1 for the homozygotes (i.e., MM and mm), respectively. Denote y_{ij} as the i^{th} transcript measurement in the j^{th} individual. An ELP locus in a marker interval can be tested, and its additive effect and interaction with the treatment conditions estimated. In ELP interval mapping, only one genetic determinant is assumed to be associated with ELPs. A putative ELP determinant can be identified by testing its additive effect and interaction with the treatment conditions, based on the following statistical model:

$$y_{ij} = \beta_0 + \beta_0^* x_j^* + \beta_1 t_i + \beta_2 g_i + \beta_1^* x_j^* t_i + \sum_{r=1}^{2R} z_{ri} \alpha_{rj} + \epsilon_{ij}$$

where $i = 1, 2, \dots, 4R$, $j = 1, 2, \dots, n$, and R is the number of array replicates per combination of treatment condition and individual; y_{ij} is the i^{th} expression level in the j^{th} individual; x^* is the indicator that the allele at the putative ELP determinant is from one of the two parental lines; t is an indicator for a treatment condition (e.g., $t_i = 0$ for control; $t_i = 1$ for treatment), and a gene (e.g., $g_i = 0$ for the structural gene A ; $g_i = 1$ for the structural gene B), respectively; z_{ri} is an indicator for the array r which was used to measure the expression level i for each individual; β_0 is the overall mean effect; β_0^* is the (additive) genotype effect of the putative ELP determinant; β_1 is the treatment effect; β_2 is the gene effect; β_1^* is the interaction between genotype of the putative ELP determinant and the treatment condition; the α 's are the array effects that are assumed to be independently and identically distributed as a random normal with mean 0 and variance σ_α^2 ; and the ϵ 's are the measurement errors distributed as a random normal with mean 0 and variance σ_ϵ^2 . Furthermore, it is assumed that the measurement errors are independent among and within individuals, and independent of the array effects.

2.2 Maximum Likelihood Estimation

If the genotype of the ELP determinant is known, the expression levels (y_{ij}) are distributed as two multivariate normal distributions (one for each parental contribution to the offspring). However, the genotype of the ELP determinant is unobservable, and needs to be estimated through the genotypic information and estimated genetic distance between the markers that

define the interval. In short, the distribution of the expression levels is described as a mixture of the two parental multivariate normal distributions.

Relying on matrix notation, the likelihood function for the parameters $\theta = (\rho, \beta_1^*, \beta_2, \sigma_\alpha^2, \sigma_\epsilon^2)$ is $L(\theta|\mathbf{Y}) = \prod_{j=1}^n p_{0j}\phi(\mathbf{y}_j; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}) + p_{1j}\phi(\mathbf{y}_j; \boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ where $\mathbf{Y} = (y_1, \dots, y_n)$, ϕ represents a standard multivariate normal probability density function, and $\rho = \frac{r_1}{r}$ is the ELP position with r_1 and r denoting the recombination frequencies between the ELP determinant and the left flanking marker and between the two flanking markers, respectively. By treating the unobservable genotype of the ELP determinant as the missing data, the maximum likelihood estimates are gained from the expectation conditional maximization (ECM) algorithm (Meng and Rubin, 1993).

2.3 Hypothesis testing

To identify an ELP determinant with an additive genotype effect, as well as interaction with the treatment condition, hypotheses are developed and the corresponding models are reduced from the previously stated full model. The hypotheses are introduced for the purpose of testing the joint genetic effect and the ELP \times treatment interaction. For statistical significance of the joint genetic effect, an empirical threshold is obtained by estimating the null distribution of maximum LRT statistics based on permutations (Churchill and Doerge, 1994) with an experiment-wise error rate of 0.05. In the same manner, a permutation-based threshold is constructed to declare the significance of interaction between the ELP determinant and treatment conditions.

3 Simulation Studies and Real Data Analysis

Simulations are conducted to compare the power of ELP interval mapping (EIM) with the existing QTL mapping methods (multiple trait interval mapping; MTIM and multiple trait composite interval mapping; MTCIM), as applied to the ELP data gained from two structural genes. The effects of the sample size and the number of array replicates on detecting genetic determinants of ELPs are also investigated. Without presenting the details, when detecting an ELP determinant and its interaction with the treatment, the EIM method showed higher power than the two existing QTL methods (across various combinations of the parameter settings, sample sizes, and number of array replicates). These results are expected since the EIM is tailored to dissect all sources of variation that define an ELP, and to overcome the limitations of the existing QTL mapping methods. The power difference in detecting an ELP determinant between the mapping methods

was larger with a small sample size ($n = 100, 200, 300$) than with a large sample size.

EIM, MTIM, and MTCIM were employed to study quantitative variation in gene expression levels (ELPs) as related to disease resistance responses in Arabidopsis. A population of 211 recombinant inbred lines (RILs), derived from two genetically distant Arabidopsis accessions, Bayreuth (Bay-0) and Shahdara (Sha) (Loudet et al., 2002), was employed. For each of 211 RILs, transcript levels of 22,810 genes/features under two treatment conditions (control vs. salicylic acid (SA) induction) were measured with two array replicates using commercial oligonucleotide arrays (Affymetrix ATH1 GeneChip microarrays; Affymetrix, 2002). After quantile normalization over the two array replicates per combination of RIL and a treatment condition, the transcript level data was natural log transformed. Based on a genetic map estimated with 540 SFP markers (West et al., 2006) and 38 microsatellite markers (Loudet et al., 2002), a 5 cM framework map of 95 selected markers (2 microsatellite and 93 SFP markers) was constructed for ELP mapping. The genotypes of the 95 markers were scored on the 211 RILs. Further experimental details are available in West et al. (2007).

Using a known gene network of 16 genes involved with SA-NPR1 regulated protein secretion (Wang et al., 2005) and the EIM analysis, ELPs for pairs of structural genes are investigated. Because these genes are in a network, some are expected to have common ELP determinants. Out of the 16 genes, 13 genes shared two ELP determinants. A total of 200 possible pairs of the 16 genes were considered, and ELPs of each pair was respectively mapped via EIM and compared to existing multiple trait QTL mapping methods.

4 Discussion

The comprehensive dissection of complex traits is an ongoing challenge. Advances in technology, namely microarrays, have allowed the investigation of complex traits at a molecular level, and when accompanied by appropriate statistical methodologies have potential to identify genetic determinants of quantitative variation in mRNA transcripts (ELPs). To date, existing statistical techniques for quantitative trait locus (QTL) mapping have been employed to detect genetic determinants of ELPs under a single condition. Due to the limitations of analyzing ELPs measured under multiple conditions (e.g., control vs. treatment), accommodating non-genetic factors, and including the treatment array effects, existing methods do a poor job of estimating the main effects and treatment interactions. Even though the proposed EIM approach provides more statistical power for testing a variety of effects, computational issues remain. As such, the EIM is limited to dissecting complex traits using pairs of genes. While identifying pairwise ELP determinants is not an optimal approach, the genes found significant

can be then used to initiate gene networks that are associated with expression level polymorphisms.

Acknowledgments: This research was supported by an National Science Foundation Arabidopsis 2010 grant (MCB-0115109) to RWD, Dina St. Clair and Richard Michelmore (UC Davis).

References

- Affymetrix (2002). *Affymetrix: Statistical algorithms reference guide*. Affymetrix, Santa Clara, CA.
- Anholt, R. R. H., and Mackay, T. F. (2004). Quantitative genetic analyses of complex behaviours in *Drosophila*. *Nature Reviews Genetics*, **5**, 838-849.
- Churchill, G. A., and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963-971.
- Doerge, R.W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics*, **3**, 43-52.
- Doerge, R.W., Zeng, Z.-B., and Weir, B. S. (1997). Statistical issues in the search for gene affecting quantitative in experimental populations. *Statistical Science*, **12**, 195-219.
- Falconer, D. S. (1952). The problem of environment and selection. *American Naturalist*, **86**, 293-298.
- Jansen, R. C. (1993). Interval mapping of multiple quantitative trait loci. *Genetics*, **135**, 205-211.
- Jansen, R. C. (1994). Controlling the type I and type II errors in mapping quantitative trait loci. *Genetics*, **136**, 871-881.
- Jansen, R. C. (2003). Studying complex biological systems using multifactorial perturbation. *Nature Reviews Genetics*, **4**, 145-151.
- Jansen, R. C., and Nap, J. P. (2001). Genetical genomics: the added value from segregation. *Trends in Genetics*, **17**, 388-391.
- Jiang, C., and Zeng, Z.-B. (1995). Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics*, **140**, 1111-1127.
- Kao, C. H., and Zeng, Z.-B. (2002). Modeling epistasis of quantitative trait loci using Cockerhams model. *Genetics*, **160**, 1243-1261.

- Kim, K., West, M. A. L., Michelmore, R. W., St. Clair, D. A., and Doerge, R. W. (2005). Old methods for new ideas: genetic dissection of the determinants of gene expression levels. In Gustafson, J. P., Shoemaker, R., and Snape, J. W., editors, *Genome Exploitation: Data Mining the Genome, The 23rd Volume in the Stadler Symposia*, pages 89-105. Springer, New York, NY, USA.
- Kim, K. (2007). *Statistical Issues in Mapping Genetic Determinants for Expression Level Polymorphisms*. Ph.D. Dissertation, Purdue University, West Lafayette, IN, USA.
- Lander, E. S., and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**, 185-199.
- Lipshutz, R. J., Fodor, S. P. A., Gingeras, T. R., and Lockhart, D. J. (1999). High density synthetic oligonucleotide arrays. *Nature Genetics*, **21**(supplement), 20-24.
- Loudet, O., Chaillou, S., Camilleri, C., Bouchez, D., and Daniel-Vedele, F. (2002). Bay-0 x Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in Arabidopsis. *Theoretical and Applied Genetics*, **104**, 1173-1184.
- Lynch, M., and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Inc., Sunderland, Massachusetts.
- Meng, X. L., and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, **80**, 267-278.
- Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature*, **405**, 847-856.
- Sax, K. (1923). The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics*, **8**, 552-560.
- Thoday, J. M. (1961). Location of polygenes. *Nature*, **191**, 368-370.
- Wang, D., Weaver, N. D., Kesarwani, M., and Dong, X. (2005). Induction of protein secretory pathway is required for systemic acquired resistance. *Science*, **308**, 1036-1040.
- West, M. A. L., van Leeuwen, H., Kozik, A., Kliebenstein, D. J., Doerge, R. W., St. Clair, D. A., and Michelmore, R. W. (2006). High-density haplotyping with microarray-based expression and single feature polymorphism markers in Arabidopsis. *Genome Research*, **16**, 787-795.

- West, M.A.L., Kim, K., Kliebenstein, D.J., van Leeuwen, H., Michelmore, R.W., Doerge, R.W., and St. Clair, D.A. (2007). Global eQTL mapping reveals the complex genetic architecture of transcript level variation in Arabidopsis. *Genetics*, **175**, 1441-1450.
- Zeng, Z.-B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 10972-10976.
- Zeng, Z.-B. (1994). Precision mapping of quantitative trait loci. *Genetics*, **136**, 1457-1468.

Penalized Spline Estimation and Mixed Models – A Flourishing Statistical Partnership

Göran Kauermann¹

¹ Centre for Statistics, Bielefeld University, Dep of Business Administration and Economics, POB 100131, (D) 33501 Bielefeld

Abstract: The paper describes the link between penalized spline smoothing and Mixed Models and how these two models form a practical and theoretically interesting partnership. Three particular offsprings of this partnership are discussed in detail. The first shows how to select the spline dimension for fitting a smooth function. The second contribution shows the partnership in a classification context. And finally, the estimation of a smooth density is considered.

Keywords: Penalized Spline Smoothing; Mixed Models.

1 Introduction

Since the seminal paper by Eilers and Marx (1996) the use of penalized spline smoothing, or P-spline smoothing as Eilers and Marx coined it, has become more and more popular in applied and recently also in theoretical statistics. The original idea traces back to O’Sullivan (1986) but the real breakthrough occurred with the book by Ruppert, Wand and Carroll (2003) who linked the idea of penalized spline smoothing to Mixed Models (see also Wand, 2003). The underlying principle and simple idea is as follows. An unknown smooth function is estimated by replacing the function by a high dimensional basis representation. For estimation a penalty is imposed on the spline coefficients, or to be more precisely on the variation of the spline coefficients, which induces a smooth fit. Making use of quadratic penalties and comprehending the penalty as a *priori* distribution yields a Mixed Model in the classical sense (see Searle, Casella & McCulloch, 1992). With this consolidation an interesting statistical partnership has started. One important practical and compelling benefit of this partnership is that the smoothing (or penalty) parameter plays the role of the *a priori* (inverse) variance of the spline coefficients, which can be estimated from the data using Maximum Likelihood. This is implemented for regression smoothing models in **R** (www.r-project.org) in the **Semipar** package accompanying the book of Ruppert, Wand and Carroll (2003) as well as in the newest **mgcv** package by Wood (2006), see also Ngo and Wand (2004). The practi-

cability and feasibility of penalized spline smoothing is probably one of the reasons why it has been investigated and applied in numerous papers in the recent years. A comprehensive and commendable survey of the last years' research activities in the field of penalized spline smoothing has been composed by Ruppert, Wand and Carroll (2009). We contribute to this work by pursuing a view focusing on the advantages and possibilities of linking penalized spline smoothing to Mixed Models. In particular we discuss in this paper three new developments in (a) choosing the spline dimension, (b) using penalized splines for classification and (c) estimating a density with penalized splines. The steering idea behind the marriage of penalized splines with Mixed Models is that one obtains a coherent objective function to optimize which is the marginal likelihood resulting after integrating out the random spline coefficients. The marginal likelihood itself can be used for more general purposes exemplary shown in this paper in the proposed setting (a), (b), (c) from above.

2 Choosing the Spline Dimension

As remarked before, the principal idea of penalized spline estimation is simple. This is primarily due to the practical convention that the spline basis is set up before estimation. Its dimension is considered to be fixed and small compared to the sample size and also (to a practical amount) independent of the sample size. This is or has been the main criticism towards penalized spline smoothing, originating particularly from the classical spline smoothing community. In the recent years, the asymptotic properties and how the dimension of the spline basis should grow with the sample size has been under fruitful investigation, see Hall and Opsomer (2005), Li and Ruppert (2008), Kauermann, Krivobokova & Fahrmeir (2009) and Claeskens, Krivobokova & Opsomer (2009). Though these papers shed some light on the $n \rightarrow \infty$ scenario, they yield little practical impact on how to select the number of splines for $n < \infty$. The central paper in this respect is Ruppert (2002) who gives a rule of thumb on how to select K , the dimension or number of knots of a spline basis, respectively. We argue in this line but utilize the partnership to Mixed Models. In fact, let y be the response, which is for the purpose of presentation assumed to be normally distributed with mean $m(x)$ and homoscedastic residual error ϵ , where $m(\cdot)$ is an unknown smooth function and x a metrical covariate. We replace $m(x)$ by $B(x)b$ with $B(x)$ as K dimensional spline basis located at knots τ_1, \dots, τ_K , say. Treating b as parameter vector we impose the penalty $\lambda b^T \tilde{D} b$, where \tilde{D} is an appropriately chosen penalty matrix. A convenient choice for $B(\cdot)$ is to use B-splines (see de Boor, 1972) and to penalize the variation of coefficients b by taking differences of neighbouring spline coefficients (see Eilers and Marx, 1996). Wand and Ormerod (2008) show that this (and other spline settings) can be rewritten to $B(x)b = X(x)\beta + Z(x)u$ with bases

matrices $X(\cdot)$ and $Z(\cdot)$ resulting by simple matrix algebra. The penalty term $b^T \tilde{D}b$ is then equivalently formulated on coefficients u only in the form $u^T D u$, where D is now of full rank. Comprehending the quadratic penalty as (proper) *a priori* distribution allows to derive the likelihood for independent observations (x_i, y_i) , $i = 1, \dots, n$ from the Mixed Model

$$Y|u \sim N(X\beta + Zu, \sigma_\epsilon^2 I_n), \quad u \sim N(0, \sigma_u^2 D) \quad (1)$$

where $Y = (y_1, \dots, y_n)^T$ and X and Z are matrices with rows $X(x_i)$ and $Z(x_i)$, respectively. We denote the likelihood resulting from (1) by $l(\beta, \sigma_\epsilon^2, \sigma_u^2)$. This likelihood is also called the *marginal likelihood* since it results by integrating out the random spline effects in (1). It should also be clear that the likelihood does also depend on the spline dimension K , which we take into account by writing $l_K(\beta, \sigma_\epsilon^2, \sigma_u^2)$. We could now consider K as additional discrete valued parameter which needs to be optimized. That is we intend to maximize $l_K(\hat{\beta}, \hat{\sigma}_\epsilon^2, \hat{\sigma}_u^2)$ for different values of K , where $\hat{\beta}$, $\hat{\sigma}_\epsilon^2$ and $\hat{\sigma}_u^2$ are the Maximum Likelihood estimates for fixed K .

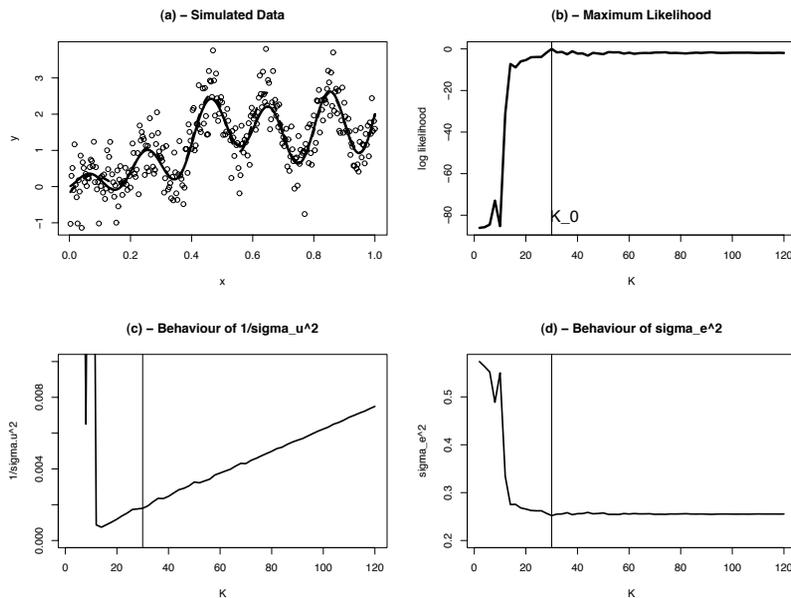


FIGURE 1. Simulated Data fitted with penalized splines (a). Value of the maximized marginal likelihood for different values of K (b). Estimate $\hat{\sigma}_u^{-2}$ (c) and $\hat{\sigma}_\epsilon^2$ (d) for different values of K .

In Figure 1 plot (b) we show the behaviour of $l_K(\hat{\beta}, \hat{\sigma}_\epsilon^2, \hat{\sigma}_u^2)$ as a function

of K . Plot (a) shows the function we simulated from and plot (c) and (d) show how the Maximum Likelihood estimates $1/\hat{\sigma}_u^2$ and $\hat{\sigma}_\epsilon^2$ change with K . The pattern seen in plot (b) is the same for a wide range of functions we simulated from (not reported here). Once the likelihood reaches its plateau, its value does not change by increasing K and the position of the maximum, indicated at K_0 in the plot, occurs for a relatively small K . In fact, it can be shown that

$$l_K(\hat{\beta}, \hat{\sigma}_\epsilon^2, \hat{\sigma}_u^2) = O_p(1 + n/K), \quad (2)$$

that is the value of the maximized likelihood does hardly change, once K is large enough. This implies practically that a relatively small K provides the optimal fit expressed in the maximum of the marginal likelihood. Numerically one can therefore start with a small K and increase K until the value of the maximum likelihood does not change. More details are found in Kauermann and Opsomer (2009). The findings can also be extended to generalized response models with non-normally distributed response variable.

Data Example

We consider data on the heating demand in the district heating network in the city of Wuppertal in North-Western Germany. Let y be the total heating demand at day i . Note that heating demand here refers to both, heating of houses as well as providing hot water. Heating is induced by steam taken from the district's steam network which is fed by two power plants in the city of Wuppertal. Steam (i.e. heating) has to be provided throughout the whole year and not only in winter months. The heating demand is modelled to depend on the continuous covariates day of the year (`ydayi`), the maximum and minimum temperature of the day (`max.tempi`, `min.tempi`) and the mean global radiation (`mean.radi`). Moreover, the heating demand depends on the day of the week (`wdayi`), where we group the days into the four categories Sunday, Monday, Tuesday to Friday and Saturday, respectively. Public holidays are generally classified as Sundays and in leap years we omit the last day of the year for simplicity. We model the heating demand as

$$y_i = \text{wday}_i + m_1(\text{yday}_i) + m_2(\text{max.temp}_i) + m_3(\text{min.temp}_i) + m_4(\text{mean.rad}_i) + \epsilon_i \quad (3)$$

for $i = 1, \dots, n$. Our data base contains the days from January 1st 2006 to December 31st 2008. Since the data are observed sequentially, it is plausible to assume that the residuals ϵ_i are serially correlated. We assume an AR(1) structure. Note that in case of correlated residuals smoothing parameter selection becomes unstable, see Opsomer, Wang & Yang (2001). The use of a Mixed Model for smoothing parameter selection however exhibits some robustness with respect to misspecification of the correlation structure, as shown in Krivobokova and Kauermann (2007). Based on this result we feel confident that even if the assumed AR(1) structure is too simplistic, the

TABLE 1. Performance of the selection algorithm of K in the heating demand example

step	K_1	K_2	K_3	K_4	log likelihood
0	4	4	4	4	-6970.949
1	6	4	4	4	-6953.221
2	6	6	4	4	-6940.536
3	6	6	4	6	-6940.433
4	6	8	4	6	-6936.722
5	6	10	4	6	-6936.516

selection of the smoothing parameter based on the Maximum Likelihood estimate in the Mixed Model will work properly. Apparently, in order maximize the marginal likelihood resulting from the Mixed Model version of (3), i.e. after replacing the unknown functions by K dimensional bases and imposing a normal prior on the spline coefficients, we need to find the optimum for a 4 dimensional vector K , each element corresponding to the dimension of the spline basis for one of the four functions. Instead of running a 4 dimensional optimization, we start with a small K and increase K sequentially over the 4 functions until the marginal likelihood does not any longer increase. We thereby increase K by a step of size 2 and Table 1 shows the outcome of the selection routine applied to the data. If the marginal likelihood does not increase for one function, we sequentially select the next function to increase K . If the likelihood does not increase for all functions when increasing K , the algorithm terminates. The resulting optimal fit based on penalized B-splines is shown in Figure 2. We see that even a small number of splines provide a satisfactory fit. The interpretation of the functions is straight forward showing an increased heating demand in winter and for cold temperatures. The effect of mean radiation is overall weak.

3 Classification with Mixed Models

In classification the task is to predict a discrete valued variable y given a set of potential classifier variables x . In the simplest scenario, variable y is binary, indicating two groups of observations and x may be metrical or categorical. As example we later make use of the spam email data set provided by Hastie, Tibshirani and Friedman (2001). Here y indicates whether an email is spam ($y = 1$) or not ($y = 0$) and x is a high dimensional vector with each component giving the percentage of particular word or character combinations in the email. The intention is to predict \hat{y} , that is to classify an incoming email as spam or not spam based on quantities x . The field of classification is thoroughly well developed with numerous successful and

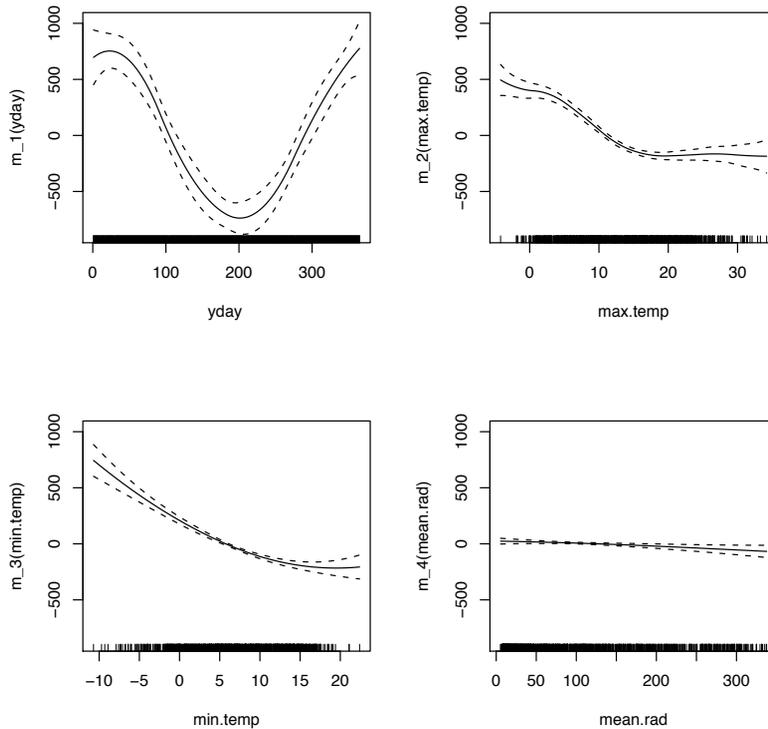


FIGURE 2. Additive model fit with optimized spline dimension showing the influence of `yday`, `max.temp`, `min.temp` and `mean.rad`.

competing methods, such as linear or quadratic discriminant analysis, neural networks, support vector machines, classification trees, to just mention a few. A detailed overview is provided in Hastie, Tibshirani & Friedman (2001). We contribute to this field by rewriting the classification problem as (additive) regression model. We suggest to pursue a model selection in a Mixed Model framework to maintain parsimony and numerical stability. Assume that vector $x = (x_1, \dots, x_p)$ contains p metrical covariates and assume that we also have q factorial (here binary) explanatory variables $v = (v_1, \dots, v_q)$, say. A full additive model for the probability of $y = 1$ would write as

$$\text{logit}\{P(y = 1|x, v)\} = \beta_0 + \sum_{j=1}^p m_j(x_j) + \sum_{j=1}^q v_j \beta_{v_j} \quad (4)$$

where $m_j(\cdot)$ are smooth but unknown functions. Note that we need additional constraints on $m_j(\cdot)$ to achieve identifiability (see Hastie and Tibshirani, 1990) which are omitted here and subsequently for ease of presentation. Replacing the unknown functions by a high dimensional basis allows to replace (4) by

$$\text{logit}\{P(y = 1|x, v)\} = \beta_0 + \sum_{j=1}^p X(x_j)\beta_{xj} + \sum_{j=1}^p Z(x_j)u_j + \sum_{j=1}^q v_j\beta_{vj}. \quad (5)$$

We fit the model after imposing a penalty on coefficients u_j in the form

$$u_j \sim N(0, \sigma_j^2 D_j). \quad (6)$$

Note that with (5) and (6) we have constructed a Generalized Linear Mixed Model (GLMM). There are now two regularizations necessary to make use of (5) in practice. First, spurious coefficients β_{xj} and β_{vj} need to be taken out of the model by setting $\beta_{xj} \equiv 0$ or $\beta_{vj} \equiv 0$. In the same way we need to set $u_j = 0$ if there is no evidence for a functional effect $m_j(x_j)$ in the data. While the first task can be handled in a classical parametric style, e.g. looking at p-values or information criteria, the second task is handled by setting $\sigma_j^2 \equiv 0$ to impose $u_j \equiv 0$. This suggests to select the model in coherent way by running a forward selection in the following style. The parameters of the full model are $\theta = (\beta_{x1}, \dots, \beta_{xp}, \sigma_1^2, \dots, \sigma_p^2, \beta_{v1}, \dots, \beta_{vq})$. One starts now with the null model by setting all parameters to zero and we denote this parameter as $\theta^{(0)}$. Letting $\theta^{(t)}$ denote the parameter after the t -th step in the iteration we calculate for the j -th component of θ with $\theta_j^{(t)} = 0$ the score

$$U_j(\theta^{(t)}) = \left. \frac{\partial l(\theta)}{\partial \theta_j} \right|_{\theta=\theta^{(t)}} \quad (7)$$

where $l(\theta)$ is the *marginal* likelihood, that is after integrating out coefficients u_j . In practice, since the true marginal likelihood is not analytical, we use a simple Laplace approximation in the style of Breslow and Clayton (1993). Our proposal is to use $U_j(\theta^{(t)})$ as selection criterion for the potential parameters in the model. If j refers to an index relating to β_x or β_v then high absolute values of $U_j(\theta^{(t)})$ (assuming standardized covariates) indicate that component θ_j should be in the model. Similarly, if j refers to a component out of $\sigma_x^2 = (\sigma_1^2, \dots, \sigma_p^2)$ then $U_j(\theta^{(t)}) < 0$ suggest that $\sigma_j^2 = 0$ while large positive values of $U_j(\theta^{(t)})$ proposes to allow for a smooth effect of variable x_j in the model. It can be shown (see Kauermann, Ormerod & Wand, 2009) that (7) is easily calculated since it is either a standard Wald statistic for index j referring to β_x or β_v or for index j referring to $\theta_j = \sigma_j^2$ one gets

$$U_j(\theta^{(t)}) = -\frac{1}{2} \text{tr}(Z_j^T \hat{W}^{(t)} Z_j D_j^{-1}) + \hat{\epsilon}^{(t)T} Z_j^T D_j^{-1} Z_j \hat{\epsilon}^{(t)} \quad (8)$$

where $\hat{\epsilon}^{(t)}$ is the fitted residual vector based on parameter estimate $\hat{\theta}$ and $\hat{W}^{(t)}$ is the diagonal weight matrix containing binomial variances. We can now successively include the covariates or smooth functions dependent on the absolute (for β_x and β_v) or positive (for σ_x^2) values of $U_j(\theta^{(t)})$. After inclusion of a component we check with an information criterion whether the component should in fact be in or not. To maintain coherence, we propose to make use of the marginal Akaike criterion suggested in Wager, Vaida & Kauermann (2007), see also Vaida & Blanchard (2005). This is defined as

$$\text{mAIC}(\hat{\theta}^{(t)}) = -2 l(\hat{\theta}^{(t)}) + 2|\hat{\theta}^{(t)}| \quad (9)$$

with $|\hat{\theta}^{(t)}|$ referring to the number of elements not set to zero. Hence smooth and parametric components are coherently penalized by its numbers of parameters in the marginal likelihood.

The procedure attracts by its coherent style, available due to the partnership between penalized spline smoothing and Mixed Models. Moreover, and possibly more importantly, the procedure performs promisingly well in practice when being compared to available routines like standard generalized additive models or BRUTO (see Hastie and Tibshirani, 1990). It beats these methods, both, in computing time and prediction error, details are provided in Kauermann, Ormerod & Wand (2009).

Data Example

We demonstrate the use of the algorithm with a classical data example in the field of classification. We make use of the ‘spam’ dataset (see Hastie, Tibshirani & Friedman, 2001) which contains data on 4601 emails with 57 metrically scaled potential predictor variables. We could, in principle, fit a generalized additive model for all 57 variables using the `gam(.)` procedure in **R** (see Wood, 2006). Alternatively, we can use the suggested forward selection routine combined with the marginal Akaike criterion. Even though the latter is a stepwise routine, it reduces the computing time compared to fitting a generalized additive model with all 57 covariates to about 1/20. We select 36 covariates out of the 57 and by doing so we can also reduce the classification error from 5.89% for the full additive model to 5.38% for our selected model. The fitted curves $m_j(x_j)$ are shown in Figure 3. More details and studies about the performance of the routine are provided in Kauermann, Ormerod & Wand (2009).

4 Penalized Density Estimation

Density estimation with penalized splines has been proposed by Silverman (1982) and has been further developed by Gu (1993) and Gu and Wang (2003). The idea is strongly related to the suggestion of Eilers and Marx (1996) who rephrase density estimation to a regression framework by binning the data, see also Ruppert, Wand & Carroll (2003). A different line

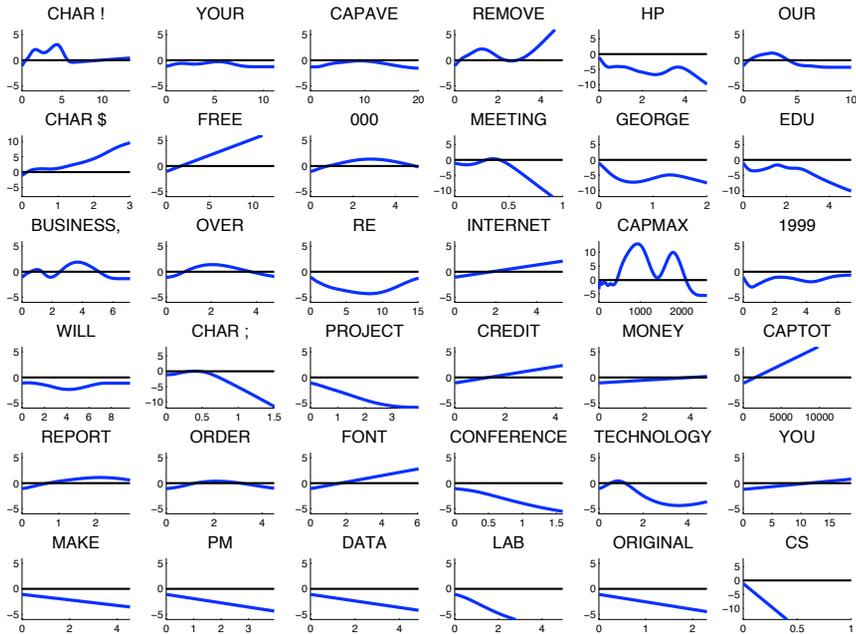


FIGURE 3. Effect of percentage of occurrence of words on classification of emails in spam or non-spam.

of penalized density estimation has been proposed in Komarek and Lesaffre (2007, 2008), where density estimation is carried out with a mixture approach. This approach will subsequently be further developed and get married with Mixed Models. We assume that response variable y has unknown density $f(y)$ which is approximated by the mixture

$$f_K(y) = \sum_{k=-K}^K c_k \phi_k(y). \quad (10)$$

Here, $\phi_k(y)$ are known densities, located at knots $\tau_{-K} \dots, \tau_K$, say, and weights c_k sum up to one so that $\int f_K(y) dy = 1$ is provided. We will call $\phi_k(y)$ subsequently basis densities. The basis densities are thereby known and fixed density functions with specified parameters. We assume that $\phi_k(y)$ is continuous on its support and converges to zero at the boundary of the support. A possible choice for the basis densities is to take $\phi_k(y)$

as Gaussian density with fixed mean μ_k and variance σ_k^2 , where the mean values μ_k may be called the knots of the basis $k = -K, \dots, K$. Numerically more stable and theoretically more appealing are B-spline densities which are standard B-splines (see de Boor, 1978) normed to be densities. We will subsequently notate the knots at which the basis densities are located as τ_k with k running from $-K$ to K for convenience. We assume, that the knots τ_k cover the range of observed values of y and their location is fixed. A typical and simple setting is to have equidistant knots which is assumed subsequently. Apparently, the number of knots plays an important role in terms of bias and variance and a small number K will lead to biased estimates while for large values of K the estimates will be wiggled. We will therefore pick up the idea of penalized spline smoothing by choosing the number of knots K in a lavish and generous way and impose a penalty to achieve smoothness. The penalty is imposed on the basis weights c_k by penalizing the variation of c_k over k .

Before getting more explicit we derive the underlying likelihood. First, to have weights c_k in (10) summing up to one we set

$$c_k(\alpha) = \frac{\exp(\alpha_k)}{\sum_{k=-K}^K \exp(\alpha_k)} \quad (11)$$

with $\alpha_0 \equiv 0$ and $\alpha = (\alpha_{-K}, \dots, \alpha_K)$. Assuming independent observations y_i , $i = 1, \dots, n$, the log likelihood takes the form

$$\alpha = \sum_{i=1}^n \left[\log \sum_{k=-K}^K c_k(\alpha) \phi_k(y_i) \right]. \quad (12)$$

Following the original idea of penalized splines we want the variation of neighbouring weights c_k to be moderate. This holds if α_k does not differ abruptly from α_{k-1} and α_{k+1} , respectively. This can be written as penalty term which is added to the likelihood yielding the penalized log likelihood

$$l_p(\alpha, \lambda) = l(\alpha) - \frac{1}{2} \lambda \alpha^T D \alpha \quad (13)$$

with $l(\alpha)$ as defined in (12) and D as penalty matrix. The next step is to comprehend the penalty as a *priori* distribution in the style of a Mixed Model. To do so we set

$$\alpha \sim N(0, \lambda^{-1} D^-) \quad (14)$$

where D^- denotes the generalized inverse of D . Apparently, the prior (14) is degenerated, which is easily corrected as follows. We decompose α into the two components u and β , respectively, such that u is a normally distributed random vector with non degenerated variance and β are the remaining components treated as parameters, see also Wand and Omerod (2008). In fact based on a singular value decomposition we have $D = \tilde{U} \tilde{\Lambda} \tilde{U}^T$ with $\tilde{\Lambda}$

TABLE 2. Mean Squared Error, in brackets the standard deviation - times 10^{-3}

	MSE (sd)	Penalized Density Estimate	Binning Estimate (Eilers & Marx (1996))	Kernel Density Estimate
(a)	$n = 100$	0.960 (0.983)	1.7350 (2.395)	2.225 (3.056)
	$n = 400$	0.174 (0.187)	0.4090 (0.559)	0.724 (0.789)
(b)	$n = 100$	12.570 (18.992)	17.500 (32.360)	24.230 (50.728)
	$n = 400$	4.894 (4.230)	5.432 (5.132)	7.444 (6.49)

as diagonal matrix with positive eigenvalues and eigenvectors $\tilde{U} \in \mathbb{R}^{\tilde{p} \times \tilde{h}}$, where \tilde{p} is the number of elements in α and \tilde{h} is the rank of D . Extending \tilde{U} to an orthogonal basis by U^\perp gives $u = \tilde{U}^T \alpha$ and $\beta = U^{\perp T} \alpha$. The penalty writes now as *a priori* assumption $u \sim N(0, \lambda^{-1} \tilde{\Lambda}^{-1})$ and we get the marginal log likelihood

$$l_m(\lambda, \beta) = \log \int |\lambda \tilde{\Lambda}|^{\frac{1}{2}} \exp \{l_p(\alpha, \lambda)\} du. \quad (15)$$

The integral is approximated with a Laplace approximation and differentiation with respect to λ provides an estimate for the penalty parameter. It can be shown in simulations that penalized density estimation in the style above outperforms kernel density estimation as well as density estimation based on binning as proposed by Eilers and Marx (1996). Exemplary we show this with two densities, see Kauermann and Schellhase (2009) for more simulations. Let (a) be y drawn from a standard normal distribution, i.e. $y \sim N(0, 1)$ and (b) from a two component normal $y \sim 0.2N(-0.5, 0.5^2) + 0.8N(0.5, 0.5^2)$. Typical simulations and its fitted densities are shown in Figure 4. For comparison we also apply the binning idea of Eilers and Marx (1996) with $2K$ basis functions and a kernel density estimate, also shown in Figure 4. We now simulate $n = 100$ and $n = 400$ observations, respectively, and calculate the Mean Squared Error based on 100 simulations. The results are given in Table 2. The proposed penalized density estimate performs best, which also occurs in other simulation settings not reported here.

Data Example

To demonstrate the procedure, we fit a density to the daily returns of the Deutsche Bank AG in 2006. We fit the data with 41 ($K = 20$) basis densities and let the smoothing parameter be chosen by maximizing (15). The resulting fit is shown in Figure 5 as solid line. For comparison we also include the density estimate based on the binning idea of Eilers and Marx (1996) and a kernel density estimate, both with optimized smoothing parameters. The estimates differ slightly and, of course, it is unclear which estimate is closer to the truth. However, bearing the simulation results

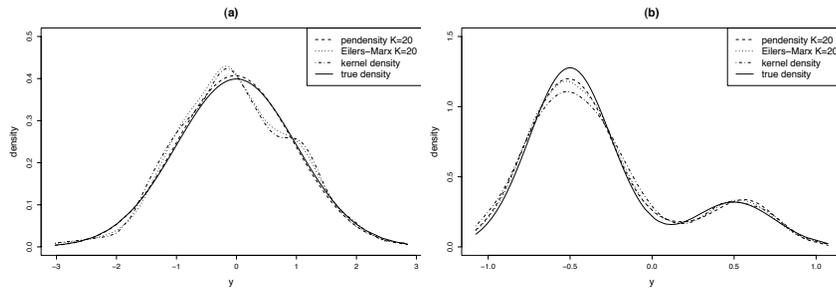


FIGURE 4. Simulation example with (a) normal density and (b) a mixture of two normal densities.

in mind there is indication that the penalized density approach proposed above is more accurate.

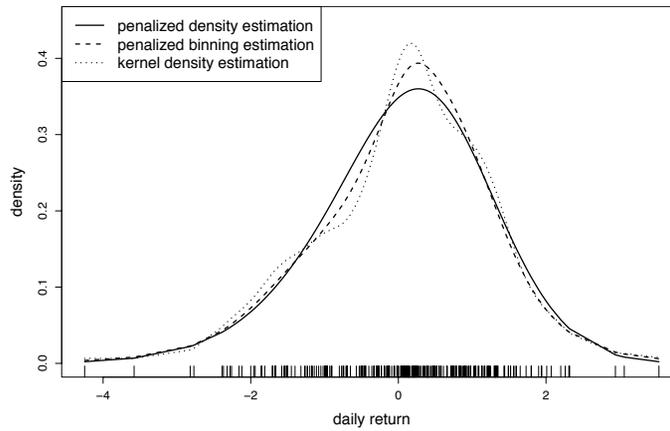


FIGURE 5. Density estimates for daily returns of Deutsche Bank AG in 2006.

5 Discussion and Extensions

The three examples described in this contribution are just a tip of an iceberg. In fact, the machinery available with linking penalized spline smoothing to Mixed Models is not fully exploited yet. The partnership opens the door to Bayesian statistics and merges the two principles to a coherent data analysis framework. In this respect one may go a step ahead on making use of penalization concepts for “normal” parameters as well, mirroring just a prior distribution on the parameter itself. It is also interesting to note that the Laplace approximation used quite centrally in the ideas discussed above deserves a better reputation than it used to have. Instead of a “poor man’s” computation for those who want to avoid computationally more complex routines like MCMC, it shows numerically simple but still accurate. This view has been also recently proposed by Rue, Martino & Chopin (2009) coming from the Bayesian world. All in all, the partnership between penalized spline smoothing and Mixed Models is in fact flourishing.

References

- Breslow, N. E., and Clayton, D.G. (1993). Approximate inference in generalized linear mixed model. *Journal of the American Statistical Association*, **88**, 9-25.
- Claeskens, G., Krivobokova, T., and Opsomer, J. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, to appear.
- de Boor, C. (1972). On calculating with B-splines. *Journal of Approximation Theory*, **6**, 50-62.
- de Boor, C. (1978). *A Practical Guide to Splines*. Berlin: Springer.
- Eilers, P.H.C., and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89-121.
- Gu, C. (1993). Smoothing spline density estimation: A dimensionless automatic algorithm. *Journal of the American Statistical Association*, **88**, 495-504.
- Gu, C., and Wang, J. (2003). Penalized likelihood density estimation: Direct cross-validation and scalable approximation. *Statistica Sinica*, **13**, 811-826.
- Hall, P., and Opsomer, J. (2005). Theory for penalised spline regression. *Biometrika*, **92**, 105-118.
- Hastie, T., and Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman & Hall.

- Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.
- Kauermann, G., Krivobokova, T., and Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society, Series B*, **71**, 487-503.
- Kauermann, G., and Opsomer, J. (2009). Data-driven selection of the spline dimension in penalized spline regression. *Technical Report*.
- Kauermann, G., Ormerod, J., and Wand, M. (2009). Parsimonious classification via generalized linear mixed models. *Journal of Classification*, to appear.
- Kauermann, G., and Schellhase, C. (2009). Penalized density estimation. *Technical Report*.
- Komarek, A., and Lesaffre, E. (2007). Bayesian accelerated failure time model for correlated interval-censored data with a normal mixture as an error distribution. *Statistica Sinica*, **17**, 549-569.
- Komarek, A., and Lesaffre, E. (2008). Generalized linear mixed model with a penalized gaussian mixture as a random-effects distribution. *Computational Statistics and Data Analysis*, **52**, 3441-3458.
- Krivobokova, T., and Kauermann, G. (2007). A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association*, **102**, 1328-1337.
- Li, Y., and Ruppert, D. (2008). On the asymptotics of penalized splines. *Biometrika*, to appear.
- Ngo, L., and Wand, M.P. (2004). Smoothing with mixed model software. *Journal of Statistical Software*, **9**, 1-54.
- Opsomer, J., Wang, Y., and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science*, **16**, 134-153.
- O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (c/r: P519-527). *Statistical Science*, **1**, 502-518.
- Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B*, **71**, 319-392.
- Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, **11**, 735-757.

- Ruppert, D., Wand, M., and Carroll, R. (2009). Semiparametric regression during 2003-2007. *Technical Report*.
- Ruppert, R., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.
- Searle, S., Casella, G., and McCulloch, C. (1992). *Variance Components*. New York: Wiley.
- Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Annals of Statistics*, **10**, 795-810.
- Vaida, F., and Blanchard, S. (2005). Conditional akaike information for mixed effects models. *Biometrika*, **92**, 351-370.
- Wager, C., Vaida, F., and Kauermann, G. (2007). Model selection for P-spline smoothing using Akaike information criteria. *Australian and New Zealand Journal of Statistics*, **49**, 173-190.
- Wand, M. (2003). Smoothing and mixed models. *Computational Statistics*, **18**, 223-249.
- Wand, M., and Ormerod, J. (2008). On semiparametric regression with O'Sullivan penalized splines. *Australian and New Zealand Journal of Statistics*, **50**, 179-198.
- Wood, S. (2006). *Generalized Additive Models*. London: Chapman & Hall.



Statistical Modeling in Oral Health Research

Emmanuel Lesaffre^{1,2}, Dominique Declerck³

¹ Department of Biostatistics, Erasmus MC, Rotterdam, Dr. Molewaterplein 50, 3015 GE Rotterdam, the Netherlands.

² Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Catholic University Leuven, Belgium and University of Hasselt, Belgium.

³ School of Dentistry, Catholic University Leuven, Belgium.

Keywords: Correlated data; Interval censoring; Misclassification; Oral Health; Random effects

1 Introduction

In dental research, more generally described as oral health (OH) research, one investigates issues on prevention, diagnosis, treatment and prognosis of diseases in the mouth, i.e. on the gums, mucosal surfaces, teeth, cranio-facial structures, etc. Consequently OH research is divided into its various sub-specialities such as preventive dentistry, cariology, endodontics, periodontology, restorative dentistry, prosthodontics, etc. yielding a variety of research questions. From a statistical perspective OH research could be considered as a particular type of medical research. However, the tooth or tooth surface is often the unit of interest even in the most basic research questions. The often highly correlated (and thus complex) nature of data in OH research therefore quickly complicates matters and requires special attention from the statistician. In this paper, we highlight some of the complexity of statistical modeling in OH research and show opportunities for statistical modelers for developing innovative approaches. No claim is made, however, that the statistical problems encountered in OH research are unique. However, in OH research the complexities meet.

Statistical issues in OH research are reviewed in this paper by means of examples of research questions in caries research which is the prime research of the second author. Examples will be primarily taken from a longitudinal oral health survey completed in 2002 in Belgium, but reference will also be made to examples from the literature.

In Section 2 a brief description of the Signal-Tandmobiel[®] (ST) survey is given. In the subsequent section some relevant dental background is provided. In Section 4 we illustrate the richness and complexity of OH research questions and indicate which statistical approaches were taken to tackle the problems. To stimulate joint research, collaboration was set up to establish joint activities and networking between the two research groups. A brief

report on activities in this sense is given in Section 5. Concluding remarks are given in Section 6.

2 Signal-Tandmobiel[®] Study

The Signal-Tandmobiel[®] survey was launched in 1996 in Flanders (North of Belgium) to obtain reliable oral health information on schoolchildren. The aims of this longitudinal study were: (1) to assess the oral health condition and its determinants in primary schoolchildren between the ages of 7 and 12 years; and (2) to implement and evaluate an oral health promotion programme offered to these children. A cohort of children was examined annually for a period of 6 years (between 1996 and 2001). At the start of the project, the children were about 7 years old. The cohort consisted of two groups: a group being examined yearly and receiving oral health education at each of these occasions (A-sample) and a (smaller) group for which an exam took place only at baseline (age 7 years) and at the end of the study (age 12 years)(B-sample). This latter group received no oral health education and served as a control for the evaluation of the impact of the intervention. In addition, in each survey year a control group of age-matched children was examined (C-samples). These children were examined only once and served as controls for cross-sectional comparisons at different ages.

Here we consider only the A-sample involving 4468 children (2315 boys and 2153 girls) born in 1989. Detailed data were collected on oral health condition using established criteria. For instance: (a) emergence stage of the permanent teeth was recorded; (b) caries experience (CE) was measured on all primary and permanent teeth up to surface level; (c) plaque scores were calculated; (d) the degree of gingivitis was established, etc. One of the aims of the study was to relate this information to the dietary and brushing behavior of the child obtained from a questionnaire filled in by the parents. Sixteen dental examiners were involved in the scoring process. In order to maintain an adequate level of reliability of the recordings and a sufficient level of agreement amongst examiners, calibration sessions were organized annually.

3 Some dental background

From an oral health perspective the children involved in the ST study were in an interesting stage of the development of their dentition. Most of the children had a mixed dentition, i.e. primary teeth co-exist in the mouth with permanent teeth which had just emerged or were about to emerge. The transition from primary to permanent dentition takes place in about 6 years.

upper right								upper left							
18	17	16	15	14	13	12	11	21	22	23	24	25	26	27	28
48	47	46	45	44	43	42	41	31	32	33	34	35	36	37	38
lower right								lower left							

FIGURE 1. FDI notation to indicate the position of a permanent tooth.

Our research interest was in factors that influence the development of caries in primary and permanent teeth and the emergence process of permanent teeth. It is popular to examine CE with a global score in the mouth summing up the CE on all primary teeth/surfaces (dmft/dmfs-score) or permanent teeth/surfaces (DMFT/DMFS-score). Such summary scores, however, do not exploit the multi-level character of the mouth. Analyses done on the unit of interest, often tooth or even tooth surface, are to be preferred.

For a good understanding of the statistical analyses it is useful to have in mind the structure of the mouth and the numbering of primary and permanent teeth in the mouth. In Figure 1 we show the FDI (Federation Dentaire Internationale) notation which indicates the position of the permanent teeth, i.e., position 24 means that the tooth is in quadrant 2 and position 4 where numbering starts from the mid-sagittal plane. For primary teeth the numbering is x1, x2, x3, x4, x5 with $x = 5, 6, 7, 8$ where for instance primary tooth 73 corresponds to the position 33 in Figure 1. Note that there are up to 20 primary teeth and 32 permanent teeth.

4 Research questions and statistical approaches

The exploration of the data collected in the ST study triggered a variety of OH research questions which lead to the application and development of novel statistical approaches. Frequentist and Bayesian approaches have been pursued.

Below we list the OH research questions and indicate which statistical approaches were invoked. It is not possible to give references for the developments as more than 50 publications (in OH and statistical) research have been realized and more than 10 manuscripts submitted or in preparation. Instead we refer to a recently published edited book on Statistical and Methodological Aspects in OH Research (Lesaffre et al. (2009)).

4.1 Initial explorations - epidemiological models

First things first: statistical analyses should always start with an exploratory part providing descriptive statistics and insight into the data(structure).

The ST data delivered valuable epidemiological information regarding prevalence and incidence of CE in Flanders. Further, the epidemiological analyses (logistic regression models) relating dietary and brushing behavior confirmed the known risk factors for CE.

4.2 Spatial structure of CE - GEE and random effects modeling

The literature indicated that there is a high degree of spatial left-right symmetry for the prevalence of CE in the mouth. However, the conclusions were based on a relatively small number of subjects. Using data from the children of the ST study at age seven and data from 3 and 5 year old children of a oral health promotion intervention study (Smile for Life Study) the spatial structure of CE in primary dentition was examined. Using GEE and latent variables models the high degree of symmetry in CE was confirmed, admitting of course that left-right equalities in prevalence cannot be proven in a statistical sense. Also the left-right association structure proved to be highly symmetrical using the Alternating Logistic Regressions approach. Further examination into the spatial structure of CE led to an apparently peculiar result that diagonally opposed teeth (say pre-molars 15 and 35) were associated in their caries experience status even when corrected for all possible confounders (e.g. caries experience status of pre-molars 25 and 45, age, gender, subject-specific random effects, etc). It turned out that this relationship disappeared on a latent score and thus was created purely by undercorrection for confounders on the manifest scale.

4.3 Geographical distribution of CE in Flanders - correction for misclassification

Exploratory analyses showed an East-West gradient in the prevalence of CE (East greater) among seven year old children. This could not be explained by geographical differences in social economic status of the regions in Flanders, nor by differences in fluoridation level in tap water. Since sixteen dental examiners were involved in scoring CE and their scoring behavior vis-a-vis a benchmark scorer also showed a East-West gradient, correction for possible misclassification was deemed necessary. A Bayesian approach was explored but also the Misclassification SIMEX approach was developed. The latter approach is an extension of the SIMEX approach introduced by Stefanski and Carroll. Correction of counts (dmft-score) and binary outcomes (CE on tooth level) for possible misclassification was considered in a cross-sectional and longitudinal context based on validation data (for which possibly corrupted and true scores are known) and employing the above mentioned approaches.

Since the validation study is most often relatively small, ways to increase its efficiency in the analysis were explored.

4.4 Optimizing the design of caries experience validation studies - determining relative efficiencies

Typically validation studies are small and not always a random sample of the main data as recommended. We showed that for progressive diseases (e.g. irreversible caries) validation data are not absolutely necessary but are useful to increase the efficiency with which parameters are estimated. Currently various designs of validation studies are under examination to explore their efficiency in combination with main data for progressive and non-progressive diseases (e.g. reversible caries). The designs take into account the multilevel structure of the dental data.

4.5 Impact of mouth, tooth, surface & examiner characteristics - multilevel modeling

Using multilevel models we are currently exploring the factors (of characteristics of surface, tooth, mouth and dental examiner) that influence the scoring behavior of the dental examiners. Multilevel modeling is also necessary to explore the longitudinal evolution of CE on tooth and tooth surface level.

4.6 Analysis of emergence of teeth and onset of caries - interval censored data

In the ST study the schoolchildren were examined annually. Hence when a new caries lesion was detected or a permanent tooth emerged, these events were only known to have taken place in-between two annual examinations. This is a particular type of censoring, called interval censoring. The non-parametric treatment of interval censoring becomes quickly complex both computationally as well as from a statistical-theoretical viewpoint. Several OH research questions necessitated to address the interval censored nature of the events. For instance, OH researchers were interested in the association of the emergence times of the permanent teeth. Further, for the analysis of the onset of caries on permanent teeth, the time at risk needs to be taken into account. Since both emergence as well as the onset of caries is interval-censored, the time at risk is doubly-interval censored and this requires a dedicated statistical approach.

The OH research questions triggered various statistical developments. For instance, when exploring the interval-censored emergence times in the ST study we experienced that the existing non-parametric computational approaches to estimate bivariate densities were hopelessly inefficient. Hence, a new algorithm for the estimation of the density of bivariate interval-censored data was developed for exploring the emergence times in the ST study. As another example, small-scale studies in the OH literature suggested that CE on primary teeth might disturb the emergence process of the

permanent teeth. To examine the dependence of the emergence process on the caries experience in primary teeth, frequentist and Bayesian approaches were developed allowing the correlation structure of the emergence times to depend on covariates (e.g. dmft-score on primary teeth).

In the presence of interval-censored data (in combination with left- and right-censored data) it is harder to claim strong parametric assumptions because of the difficulties in verifying the assumptions made. On the other hand, many of the OH research questions posed in the context of the ST study excluded the use of non-parametric approaches since they involved the impact of covariates. For this reason we explored statistical approaches that used on a flexible shape for the survival density functions, i.e. being a mixture of normal distributions on the log-scale in combination with an accelerated failure time model. This flexible distribution was assumed for the random effects distributions when survival times of several teeth (in our case molars) were examined.

Finally, we also explored the use of non-parametric Bayesian approaches to allow for flexible distributions. This is a promising Bayesian, though highly computationally involved, approach.

4.7 Future and other statistical developments in OH research

In the introduction we pointed out that the complexities meet in OH research. This can already be seen from the above sections. For another example note that in the ST study the onset of caries is interval-censored and at the same time possibly scored with error. An appropriate statistical analysis needs to tackle both problems at the same time, something we are currently pursuing.

Apart from the ST study we have been involved in other OH research projects. For instance, in a survival analysis on implants conducted in the Netherlands 5 different restoration-crown combinations were considered and up to four restorations per individual were applied. Several covariates on tooth and mouth level might be predictive for the long term success of the restoration. In this respect we are currently examining a new measure for predictive ability of the survival model in a multi-level context applied to choose the most predictive survival model for the success of the restoration.

The above enumeration of statistical developments was quite idiosyncratic. It should be seen, however, only as a reflection of our efforts to address the OH research questions in an appropriate manner thereby often requiring the development of new methodology.

No doubt, a lot of exciting statistical developments in OH research is happening at other institutions. Some of these developments can be found in the edited book of Lesaffre et al. (2009). We also draw the attention to an interesting Bayesian spatial model for periodontal disease by Reich et al. (2007).

5 Collaborative effects

Bio-statistical research resulting from research questions is rewarding in many ways. Statistical and OH research can benefit from a close collaboration: bio-statisticians are confronted with challenging statistical problems and OH researchers are less hampered by technical difficulties in their search for answers. For this reason, bi-annual international dental-statistical meetings have been organized since 2004 bringing together the two disciplines. Other initiatives have been taken by established statistical societies such as: ISCB and IBS. Most importantly these initiatives create a network of researchers of both kinds willing to exchange ideas.

6 Concluding remarks

The aim of this paper was to demonstrate that OH research is a fruitful area for the statistical modeler challenging him/her in various ways and producing statistical developments that are not only of interest for OH research but also for many other domains of medical research. For instance, our research into interval censored data is also relevant for HIV & AIDS research where the exact onset of the infection and the switch from HIV to AIDS is never exactly known.

Acknowledgements

The statistical developments were supported by Research Grants OT/00/35 and OT/05/60, Catholic University Leuven. Data collection was supported by Unilever, Belgium. The Signal-Tandmobiel[®] project comprises following partners: D. Declerck (Dental School, Catholic University Leuven), L. Martens (Dental School, University Ghent), J. Vanobbergen (Oral Health Promotion and Prevention, Flemish Dental Association), P. Bottenberg (Dental School, University Brussels), E. Lesaffre (Biostatistical Centre, Catholic University Leuven) and K. Hoppenbrouwers (Youth Health Department, Catholic University Leuven; Flemish Association for Youth Health Care). We also acknowledge support from the Interuniversity Attraction Poles Programs P5/24 and P6/03 – Belgian State – Federal Office for Scientific, Technical and Cultural Affairs.

Finally, the above described research was done in collaboration with many OH and statistical researchers: Olumide Agbaje, Kris Bogaerts, Silvia Cecere, Maria-Jos Garcia-Zattera, Alejandro Jara, Arnošt Komárek, Helmut Küchenhoff, Roos Leroy, Samuel Mwalili, Timothy Mutsvari, David Todem, Jacky Vanobbergen and Robin Van Oirbeek.

References

- Lesaffre, E., Feine, J., Leroux, B. and Declerck, D. (2009) *Statistical and Methodological Aspects of Oral Health Research*. Wiley-Blackwell, 2009 (ISBN 13: 9780470517925)
- Reich, B.J., Hodges, J.S. and Carlin, B.P. (2007) *Spatial Analyses of Periodontal Data Using Conditionally Autoregressive Priors Having Two Classes of Neighbor Relations*. *Journal of the American Statistical Association*, 102: 44-55

Part 2
Contributed papers



Hierarchical space-time models for fire ignition and percentage of land burned by wildfires

M.A. Amaral-Turkman¹ and K.F. Turkman²

¹ CEAUL-FCUL, Bloco C6, Campo Grande, 1749-0116 Lisbon, Portugal

Abstract: Policy responses for local and global fire management as well as international green-gas inventories depend heavily on the proper understanding of the annual fire extend as well as its spatial variation across any given study area. Proper statistical models are important tools in quantifying these fire risks. We propose a Bayesian method to model jointly the probability of ignition and fire sizes in Australia. The data set on which we base our model and results consists of annual observations of several meteorological and topological explanatory variables, together with the percentage of land burned over a grid with resolution of 1 degrees across Australia and New Zealand. Our model and conclusions bring improvements on the results reported by Russell-Smith *et al.* (2007) based on a similar data set.

Keywords: Wildfires; Hierarchical models.

1 Introduction

Wildland fires cause extensive loss of property and life and inflict heavy damage on ecosystems in many parts of the world every year. The role of wildfires in greenhouse gas emissions and therefore on the climate change scenarios is not particularly well understood. Ignition risk as well as the extend of wildfires depend on complex combination of climatological factors and flora to large extend, and many other factors such as human factor to smaller extend.

Australia is particularly prone to wildfires, primarily due to the devastating combination of climatological factors and flora, and lesser extend, due to many other observed and unobserved factors acting on various spatial scales. Together with the recent advances in satellite imagery, many quantitative assessment of the spatio-temporal variation of fire risk using statistical models on continental scale have been made. We refer to Russell-Smith *et al.* (2007) for the most recent and extensive study. We will take this study as the bench mark, with which we will compare our methods and conclusions, therefore we give their results in detail. In Russell-Smith's study, Australian active fire detection and mapping data sets were used.

Their data is defined on a continental grid of $0.5^\circ \times 0.5^\circ$ cells, totalling 3026 grid points. Fire affected areas in each grid between 1997-2004, derived from satellite imagery was used as the response variable, whereas rainfall, lighting, NDVI(Normalized Difference Vegetation index), Vegetation type, Fuel type, land elevation, surface roughness, land use and cadastral density observed in each grid cell over the same period, used as the set of exploratory variables. No attempt was made to model the temporal and spatial variation in these exploratory variables. The objective of this study is to explore the relative importance of the exploratory variables in study and as a consequence, assess the spatial variation of fire affected areas in a continental scale. In order to achieve these objectives, Russell-Smith *et al.* (2007) use generalized linear modeling to analyze three different response variables:

- Arcsine transformation of the annual proportion of land burned in each grid cell is regressed on the explanatory variables, using standard regression method, namely assuming Gaussian distribution for the independent error terms.
- Frequency with which any portion of a grid had fire, assuming binomial distribution and using a logit link.
- Grids with at least one fire during the study period versus grids without any fire, again assuming binomial error distribution and logit link

The latter two GLM are carried to study the spatial variation of ignition probability as well as the impact of the exploratory variables, whereas the former GLM is carried to analyze the spatial variation of percentage of land burned.

In this paper, we revisit the same problem but with a different data set and a different methodological approach. Our objectives will be same, namely infer on the spatial and temporal variation of the ignition probability as well as the percentage of land burned, given that there is ignition, over all Australia and New Zealand. The novelty will be on the statistical methods and techniques used, which we believe are substantially more informative. The data set we have is the percentage of land burned obtained by satellite imagery, defined over a grid with resolution $1^\circ \times 1^\circ$ across the whole world, obtained annually over the years 1998-2006, together with a set of exploratory variables defined over the same grid structure and time. Although the data set we use neither have the same resolution of the data nor the same exploratory variables used by Russell-Smith *et al.*(2007), there is scope for comparing the results globally. Hence, for comparative reasons, we will restrict our the study to Australia and New Zealand.

We believe that the statistical methods used by Russell-Smith can be improved in many directions to obtain better results:

1. Percentage of land burned as well as the occurrence of ignition in each grid cell show high degree of spatial and temporal dependency

for low resolution grid data, and this dependence should exist even more stronger in the high resolution data use by Russell-Smith *et al.* (2007). Although part of this dependence is explained through the exploratory variables, significant portion of this dependence is due to latent, unaccounted spatially and temporarily varying variables. Therefore standard Gaussian regression based on independent errors may not be correct. This unaccounted dependence structures have particularly undesired and strong effects on the statistical tests carried on the regression parameters.

2. Our data set contains unusual number of 0 values (zero-inflated), that is, there are many grid points where there is no ignition during the year, and the arcsine transformation is not particularly good if a substantial number of the proportions are equal to 0 or 1 or for values at the extreme ends of the possible range of p (near 0 and near 1). Here we considered a non-standard heavy tailed Student distribution for a logit transformation of the positive percentage of land burned.
3. Latent, unaccounted spatially and temporarily varying variables add significant bias to the residuals and result in the underestimation of the structural variation in the response variables. Therefore, inclusion of latent, spatially and temporarily colored random factors in the link functions and in the regression should reduce bias and improve the overall quality of the regression. They may also indicate future research directions in looking for other exploratory variables having direct effect on fire regimes.
4. Russell-Smith *et al.* (2007) suggest making inference on the ignition probability and percentage of land burned in two separate studies. However, these two events are interconnected and the inference should in principle be made jointly within one model.
5. Finally, the use of Bayesian hierarchical models brings with it the usual benefits of Bayesian methodology, such as incorporating sampling variation of model parameters as well as prior knowledge of the experts in the model.

2 Data

All data surfaces used for analysis were compiled on a grid of $1.0^\circ \times 1.0^\circ$ cells covering Australian and New Zealand land mass, resulting in an equidistant lattice with 750 cells. This is a relatively coarse grid as compared to the grid used by Russell-Smith *et al.* (2007). The region is divided into pixels of 1 degree, resulting in a lattice with 750 cells. Let $s_j = (\textit{latitude}, \textit{longitude})$ of the j -th pixel centroid in this lattice. Let $Y(s_j, t)$ denote the percentage

of land burned in pixel s_j , during year $t = 1998, \dots, 2006$. The following covariates were observed in each pixel:

1. X_1 (id): A line identifier (1-750 for 1998 ; 1-6750 for 1999-2006, from the repeated 750 grid cells for each of the 9 years)
2. X_2 (lat): latitude of the middle of each 1 grid cell
3. X_3 (lon): longitude of the middle of each 1 grid cell
4. X_4 : (rtdry): Dry Season Severity from TRMM satellite data (native resolution 0.25, aggregated to 1)
5. X_5 : (rtrp): Precipitation over the wet season (or growing season) from TRMM satellite data
6. X_6 (hf): Human footprint (static)
7. X_8 (glct): Tree landcover (from 0 to 1, 1 being 100%)(static)
8. X_9 : (glcgs): Grass and Shrubs landcover (static)
9. X_{10} (glca): Agricultural Landcover (static)
10. X_{11} (cmapp): Maximum number of consecutive dry pentads (pentad=5 days) over the corresponding year.
11. Y : (y): Estimated Burned Fraction over the year (dependent variable)

Hence, there are $n = 750$ locations, each having $N = 9$ repetitions from 1998 to 2006. Some of these covariates are temporarily static. There are other independent variables which are known to be very influential on fire regimes and sizes, such as heterogeneity in the vegetation, wind speed and direction, local topology, etc., which are not included in the analysis due to unavailability of data. These unobserved covariates will be represented in the model by a latent, spatially and temporarily colored random effect. Some of the covariates in the list exhibit strong time and space variation, but at this stage, there will be no attempt to model these temporal and spatial structures in these independent variables; they will simply be included in our model as fixed explanatory variables.

3 Acknowledgements

This work was partially financed by FCT project PTDC/MAT/64353/2006

References

- Russell-Smith, J. et al (2007). *International Journal of Wildland Fires*, **16**, 361-377.

Bayesian Approaches for Random Effects Models in Microarray Analysis

Haim Y. Bar¹ and Elizabeth D. Schifano^{1,2}

¹ Department of Statistical Science, Cornell University, Ithaca, NY 14853;
<hyb2@cornell.edu>, <eds27@cornell.edu> .

² Communicating Author.

Abstract: A linear model involving a mixture distribution is considered for the comparison of microarray data from two treatment groups. Model fitting using an empirical Bayes approach has been shown to be both accurate and numerically stable. The posterior odds of treatment/gene interactions, derived from the model, involve shrinkage estimates of both the interactions and the gene-specific error variances, leading to powerful inference. We show the same model can be fit under a fully Bayesian framework, allowing increased flexibility in terms of prior distributional assumptions and posterior inference.

Keywords: EM algorithm; Empirical Bayes; MCMC; Linear Model; LEMMA

1 Motivation and Introduction

In the microarray literature, a common goal is to identify genes that are differentially expressed between two treatments. As technology improves, the number of genes (G) simultaneously assayed continues to grow while the number of samples (n) typically remains limited, thus rendering the usual gene-wise t-statistics unsatisfactory in terms of average power. The statistical community has tried to combat the so-called ‘small n , large G ’ problem by designing more powerful statistics, typically via shrinkage of either the numerator (mean effects) or denominator (error variance effects) of the t-statistic for stabilization (e.g. Lonnstedt et al., 2002; Kendzioriski et al., 2003; Wright et al., 2003; Smyth, 2004; Cui et al., 2005). Recently, Hwang et al. (2007) proposed to shrink both the mean and the variance in the F_{SS} statistic, which proved to be the most powerful in simulations. The linear model-based approach proposed in Bar et al. (2009) for comparing (normalized) microarray data from two treatment groups leads to six different models depending on the assumptions imposed on the gene-specific effects. Interestingly, the likelihood ratio statistics from these models correspond to the statistics cited above. In particular, the RR model, in which both the nonnull gene-specific treatment effects and gene-specific error variances are modeled as random variates, leads to James-Stein-type shrinkage estimation. The resulting likelihood ratio is similar in form to

that of Hwang et al. (2007). Specifically, the RR statistics enjoy shrinkage in both the numerator and denominator of a posterior t-statistic, resulting in powerful statistics while maintaining few false positives in simulations. Their fitting approach, LEMMA (Laplace approximated EM Microarray Analysis), yields stable and accurate parameter estimates, even for the notoriously difficult mixture parameter p_1 . We show that the same RR model can easily be fit under a fully Bayesian framework, enabling increased flexibility in terms of distributional assumptions and posterior inference, while maintaining accuracy in estimation and high average power.

1.1 Model and Notation

We follow the same model and notation as in Bar et al. (2009), which we include here for the sake of completeness.

Let y_{ijg} denote the response (e.g. log expression ratio) of gene g , for subject (replicate) j , in treatment group $i = 1, 2$. We suppose a linear model,

$$y_{ijg} = \mu + \tau_i + \gamma_g + \psi_{ig} + \epsilon_{ijg}, \quad (1)$$

and assume that the errors are distributed as

$$\epsilon_{ijg} \stackrel{iid}{\sim} N(0, \sigma_{\epsilon, g}^2) \quad (2)$$

for $j = 1, \dots, n_{ig}$, independently across genes and treatments. We impose identifiability constraints, $\tau_1 + \tau_2 = 0$, and $\psi_{1g} + \psi_{2g} = 0$ for all $g = 1, \dots, G$. Then $\tau = \tau_1 - \tau_2$ is the main effect of treatment averaged across genes, and $\psi_g = \psi_{1g} - \psi_{2g}$ are the gene-specific treatment effects.

We further suppose that each gene belongs to one of two groups, either the *null* group in which $\psi_g \equiv 0$, or the *nonnull* group in which $\psi_g \neq 0$. The primary objective is to classify genes as null/nonnull based on the observed responses. Thus, we assume each gene has prior probability p_1 of being nonnull (and $p_0 = 1 - p_1$ of being null) and use Bayes rule to determine the posterior probabilities of group status, given the data.

By sufficiency, model (1) with assumption (2) allows us to restrict attention to the sum and difference of gene-specific treatment means, respectively $s_g = \bar{y}_{1 \cdot g} + \bar{y}_{2 \cdot g}$ and $d_g = \bar{y}_{1 \cdot g} - \bar{y}_{2 \cdot g}$, and the gene-specific mean squared errors,

$$m_g = \sum_{i=1}^2 \sum_{j=1}^{n_{ig}} (y_{ijg} - \bar{y}_{i \cdot g})^2 / f_g,$$

where $f_g = n_{1g} + n_{2g} - 2$. Notice that $s_g | g \sim N[2\mu + 2\gamma_g, \sigma_g^2]$, where $\sigma_g^2 \equiv \sigma_{\epsilon, g}^2(1/n_{1g} + 1/n_{2g})$, and $|g$ denotes conditioning on any gene-specific random effects. It follows that s_g carries no information about the gene-specific treatment effect ψ_g . Indeed, the LEMMA estimation procedure uses only the marginal likelihood based on $(\{d_g\}, \{m_g\})$.

Model (1) with assumption (2) also imply that d_g and m_g are conditionally independent. Specifically, we have $d_g|g \sim (1 - b_g)N_0 + b_gN_1$ independently of $m_g|g \sim \sigma_{\epsilon,g}^2\chi_{f_g}^2/f_g$, where b_g , $g = 1, \dots, G$, denote independent Bernoulli(p_1) latent indicators of nonnull status for the G genes, N_0 and N_1 denote normal variates with unequal means τ and $\tau + \psi_g \neq \tau$ respectively, but equal variances σ_g^2 , and $\chi_{f_g}^2$ denotes a chisquared variate with f_g degrees of freedom.

As in Bar et al. (2009), we choose to model all gene-specific effects as random variates, leading to the so-called ‘RR model’ (**R**andom gene-specific treatment and **R**andom gene-specific error variance effects). That is, we assume $\gamma_g \stackrel{iid}{\sim} N(0, \sigma_\gamma^2)$, $\sigma_{\epsilon,g}^{-2} \stackrel{iid}{\sim} \text{Gam}(\alpha, \beta)$, and for nonnull genes, $\psi_g \stackrel{iid}{\sim} N(\psi, \sigma_\psi^2)$.

1.2 Empirical Bayes Approach - LEMMA

Bar et al. (2009) considered an approximate EM algorithm for fitting the RR model, with gene group membership (null/nonnull), or equivalently, the latent indicators $\{b_g\}$, playing the role of the missing data. The complete data likelihood based on $(\{b_g\}, \{d_g\}, \{m_g\})$, $L_C(\phi)$, for parameters $\phi = (p_1, \tau, \psi, \sigma_\psi^2)$ in such a model requires integration over the random components, making direct application of the EM algorithm intractable. Instead, they proposed using a Laplace approximation, $\tilde{L}_C(\phi)$, to define $Q(\phi, \phi^{(m)}) = E_{\phi^{(m)}}[\log \tilde{L}_C(\phi) | \{d_g\}, \{m_g\}]$ in the $(m + 1)^{st}$ E-step, where $\phi^{(m)}$ is the current estimate of ϕ . Tractable update equations for ϕ are obtained in the $(m + 1)^{st}$ M-step when maximizing $Q(\phi, \phi^{(m)})$ with respect to ϕ . Estimates of α, β are obtained via numerical maximization of the marginal likelihood based on $\{m_g\}$. Bar et al. (2009) derived the estimated value of the likelihood ratio based on the RR model as

$$\frac{L_{0,g}}{L_{1,g}} \propto \left(\frac{\tilde{\sigma}_g^2}{\tilde{\sigma}_\psi^2 + \tilde{\sigma}_g^2} \right)^{-1/2} \exp \left\{ -\frac{1}{2} T_g^2 \right\}, \quad (3)$$

where $\tilde{\sigma}_g^2$ is an estimate of $\sigma_{\epsilon,g}^2(1/n_{1g} + 1/n_{2g})$ based on its posterior mode. The statistic T_g is a posterior t-statistic, being the ratio of the estimated posterior expectation of ψ_g to its estimated posterior standard deviation. Genes are classified as nonnull based on the posterior odds, $\hat{p}_0 L_{0,g} / \hat{p}_1 L_{1,g}$, or equivalently, the local false discovery rate (fdr) (Efron, 2005), $fdr_g = \hat{p}_0 L_{0,g} / (\hat{p}_0 L_{0,g} + \hat{p}_1 L_{1,g})$, less than a fixed cutoff.

2 Fully Bayesian Approach - MCMC

A fully Bayesian estimation approach for models such as RR, now treating all parameters as random, is simple to implement using available samplers such as WinBUGS (Spiegelhalter et al., 2003). If our interest remains in

the gene-specific treatment effect, ψ_g , we can consider building a sampler based on the observed $\{y_{ijg}\}$ directly, or, more similarly to the LEMMA approach, building a sampler based on $(\{d_g\}, \{m_g\})$. We refer to these respectively as MCMC(y) and MCMC(d). Indeed, MCMC(d) is much more computationally efficient than MCMC(y), but we show in simulation that both samplers are viable options.

For the simulations and case study in Section 3, we specified noninformative prior distributions on all parameters in the MCMC samplers except for p_1 . Since we generally expect the proportion of nonnull genes to be small, we used $p_1 \sim \text{Beta}(0.01, 0.99)$ such that the mean of the distribution is 0.01. The remaining prior distributions are listed in Table 1. In all cases, convergence was assessed using autocorrelation and trace plots.

For inference, the posterior means of the latent indicators of nonnull status, $\{b_g\}$, are used to estimate the posterior probabilities of being nonnull. Genes with null posterior probability less than a fixed cutoff are classified as nonnull.

Note that changing the distributional assumptions in LEMMA would require a new derivation of the test statistic. However, in MCMC, this would simply require modifying the distributions in the sampler, as demonstrated in the case study below.

3 Comparison of the Bayesian Approaches

We summarize the results of a simulation study, similar to that of Bar et al. (2009), to compare the performance of the two Bayesian approaches. In addition, we provide results from both Bayesian approaches applied to a case study for further comparisons.

3.1 Simulation Studies

For a range of ψ , the mean nonnull treatment effect, we generated 20 datasets according to the RR model, each with 475 null and 25 nonnull genes with parameters as in Table 1, and $\mu = 0$, $\sigma_\gamma^2 = 0.25$. In MCMC, we used prior densities as in Table 1, and $\mu \sim N(0, 100)$, $\sigma_\gamma^{-2} \sim \text{Gam}(2, 0.25)$. Note that μ and σ_γ^{-2} are not estimated in LEMMA or MCMC(d), as their effects cancel when considering the differences, $\{d_g\}$. Our samplers involved one chain of 5000 iterations, 3000 of which were burn-in. Of the remaining 2000 samples, we used a thinned 1000 samples for posterior inference/estimation.

For both Bayesian approaches, a gene was classified as nonnull if its null posterior probability (i.e., local *fdr* in the LEMMA context) was less than 0.2, as suggested by Efron (2005). We defined average power, AP , as the proportion of nonnull genes correctly classified, and the average number of false detections, AFD , as the average number of null genes incorrectly

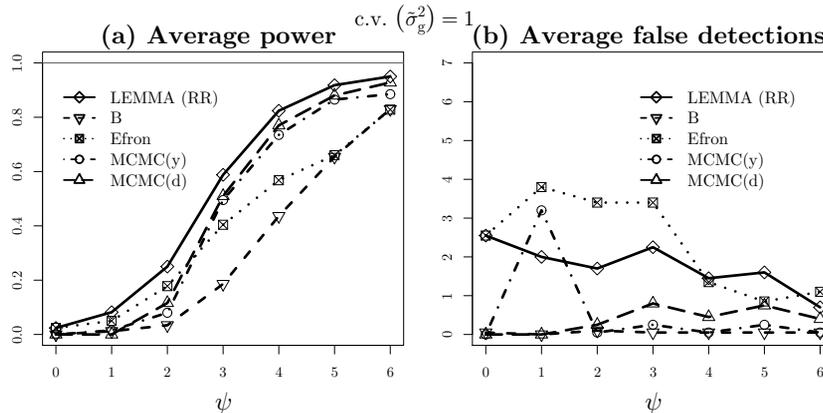


FIGURE 1. Average power and false detections with $n_1 = n_2 = 6$ and $G = 500$ using low error variance variability ($\alpha = 5$ and $\beta = 12$) under the RR model.

classified. For comparison, we also considered the B statistic (Lonnstedt et al., 2002) computed from the *LIMMA* R package (with default value $p_1 = 0.01$), and Efron’s local fdr statistics computed from the *locfdr* R package (Smyth et al., 2008; Efron et al., 2007). In Figure 1, the LEMMA statistics display the highest AP for all ψ considered. All methods yield low AFD . While MCMC does not achieve quite as high AP as LEMMA, it generally results in lower AFD .

Both LEMMA and MCMC produce accurate parameter estimates for the RR model. The estimates and parenthetical standard deviations for the case of $\psi = 4$ are provided in Table 1. Note that the LEMMA values are computed from 20 estimates, whereas those for MCMC are obtained by averaging 20 posterior means/variances (each computed from 1000 posterior samples). Not surprisingly, the LEMMA and MCMC(d) estimates are quite similar since they are both based on the data $(\{d_g\}, \{m_g\})$. We find that MCMC(y) tends to estimate p_1 more conservatively, resulting in lower AP , but also lower AFD . Although not shown, we found that as ψ increases, all methods provide more accurate estimates. For smaller values of ψ , LEMMA appears to be more robust than MCMC(y) and MCMC(d). However, for larger values of ψ , MCMC(y) and MCMC(d) tend to more accurately estimate both ψ and σ_ψ^2 , whereas LEMMA tends to underestimate ψ and compensate by overestimating σ_ψ^2 .

3.2 Case Study

We fit both approaches, specifically LEMMA and MCMC(d) with the RR model, to the publicly available, two-channel gene expression microarray

TABLE 1. Estimates for parameters at $\psi = 4$ and MCMC priors.

	TRUE	LEMMA	MCMC(y)	MCMC(d)	MCMC priors
α	5.00	5.35 (0.94)	5.71 (0.69)	5.42 (0.60)	$U(2.001, 20)$
β	12.00	13.04 (2.53)	14.61 (2.18)	13.19 (1.64)	$U(0.05, 20)$
p_1	0.05	0.07 (0.03)	0.04 (0.01)	0.07 (0.02)	$Beta(0.01, 0.99)$
τ	0.00	-0.02 (0.04)	0.01 (0.02)	-0.01 (0.05)	$N(0, 100)$
ψ	4.00	3.52 (0.88)	4.37 (0.27)	3.47 (0.57)	$N(0, 100)$
σ_ψ^2	1.00	1.80 (2.03)	1.49 (0.88)	1.18 (0.54)	$InvGam(2, 3)$

ApoA1 dataset (Callow et al., 2000). The experiment used gene targeting in embryonic stem cells to produce mice lacking apolipoprotein A-1, a gene known to play a critical role in high density lipoprotein (HDL) cholesterol levels. The responses of interest, $\{y_{ijg}\}$, are the normalized \log_2 fluorescence ratios (with respect to a common reference) where $i = 1, 2$ is the population index (control/knockout), $j = 1, \dots, 8 = n_i$ is the mouse index, and g is one of $G = 5548$ ESTs.

The parameter estimates for LEMMA and MCMC(d) under the RR model are provided in the first two columns of Table 2. Because the fully Bayesian approach allows us to easily change the distributional assumptions, we also considered MCMC(d) with a log-normal distribution on the error variances, i.e., $\sigma_{\epsilon,g}^2 \sim LN(\theta, \nu^2)$, rather than inverse gamma. The estimates obtained from this approach, labeled MCMC(d)-ln, appear in the last column of Table 2. Note that both MCMC(d) samplers used 10000 iterations, 5000 of which were burn-in. Of the remaining 5000 samples, we used a thinned 2500 samples to define our posterior distributions. Hence, estimates from both MCMC(d) samplers are posterior means based on 2500 posterior samples. Posterior standard deviations for the MCMC(d) samplers are also provided in parentheses in Table 2.

The parameter estimates corresponding to $\sigma_{\epsilon,g}^2$ are not provided in the table, as MCMC(d)-ln does not estimate α and β , but rather θ and ν^2 . The estimates for α and β in LEMMA and MCMC(d) are very similar (LEMMA: $\hat{\alpha} = 1.87$, $\hat{\beta} = 11.11$; MCMC(d): $\hat{\alpha} = 1.85(0.041)$, $\hat{\beta} = 11.14(0.32)$) yielding inverse gamma distributions with theoretical means of 0.103 and 0.106, respectively. Using the MCMC(d)-ln estimation, we obtained $\hat{\theta} = -2.74(0.012)$ and $\hat{\nu}^2 = 0.69(0.016)$, with theoretical mean of 0.091.

The prior distributions used in this case study were the same as those in the simulations with the exception of α , for which we used $U(1.001, 20)$. In the MCMC(d)-ln case, we used the following prior distributions: $\theta \sim N(0, 100)$ and $\nu^{-2} \sim Gamma(3, 2)$. We found in this study that MCMC(d) is much more computationally efficient than MCMC(y), and was less sensitive to the choice of initial values for the sampler.

TABLE 2. Estimates for ϕ in ApoA1 Case Study.

	LEMMA	MCMC(d)	MCMC(d)-ln
p_1	0.0039	0.0020 (0.001)	0.0018 (0.0007)
τ	0.0075	0.0074 (0.002)	0.0066 (0.0015)
ψ	0.69	1.12 (0.35)	1.45 (0.25)
σ_ψ^2	0.84	0.74 (0.22)	0.80 (0.19)

Using LEMMA with the 0.2 posterior probability threshold, we detected 9 nonnull genes including the ApoA1 gene and others closely related to it. The top eight genes had local *fd*r values of nearly zero, while the ninth had a much higher value of 0.08. Using either MCMC(d) or MCMC(d)-ln under the RR model with the same threshold, only the top eight of the nine genes identified in LEMMA were classified as nonnull. This same set of the top eight genes were also identified as nonnull (among others) when using the *LIMMA* and *locfdr* R packages, and were confirmed to be differentially expressed in the knockout versus the control line by an independent assay. Interestingly, assuming that only these eight genes are in the nonnull group, the true value of p_1 is 0.00144. The LEMMA estimate is 0.0039, and the MCMC(d) and MCMC(d)-ln estimates are 0.002 and 0.0018, respectively; the estimates from the *locfdr* package using the MLE and CME methods are -0.036 and -0.083 , respectively. The *LIMMA* package does not provide an estimate for p_1 , and uses a default value of 0.01 to fit the model.

3.3 Conclusions

We have demonstrated that the fully Bayesian approach is competitive with LEMMA (Bar et al., 2009). LEMMA offers slightly more powerful test statistics and faster computational time, whereas MCMC provides more flexibility in model specifications and fewer false detections. Both approaches provide accurate estimates. Additionally, MCMC provides posterior samples, as opposed to just point estimates, and in principle can easily be extended to a multivariate model by selecting appropriate priors.

References

- Bar, H.Y., Booth, J.G., Schifano, E.D., and Wells, M.T. (2009). Laplace approximated EM Microarray Analysis: an empirical Bayes approach for comparative microarray experiments. *Under Review*.
- Callow, M., Dudoit, S., Gong, E., Speed, T., and Rubin, E. (2000). Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Research*, **10**, 12.

- Cui, X., Hwang, J.T.G., Qiu, J., Blades, N., and Churchill, G. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, **6**, 59-75.
- Efron, B. (2005). Local false discovery rates.
<http://www-stat.stanford.edu/~brad/papers/False.pdf>.
- Efron, B., Turnbull, B.B., and Narasimhan, B. (2007). locfdr: Computes local false discovery rates. *R package*, version 1.1-5.
- Hwang, J.T.G., and Liu, P. (2007). Optimal tests shrinking both means and variances applicable to microarray data. *Under Review*.
- Kendzierski, C.M., Newton, M.A., Lan, H., and Gould, M.N. (2003). On parametric Empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*, **22**.
- Lonnstedt, I., and Speed, T. (2002). Replicated Microarray Data. *Statistica Sinica*, **12**, 31-46.
- Smyth, G.K. (2004). Linear Models for Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**, 1.
- Smyth, G.K., Ritchie, M., Thorne, N., and Wettenhall, J. (2008). limma: Linear Models for Microarray Data User's Guide.
<http://www.bioconductor.org>.
- Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2003). WinBUGS User Manual Version 1.4. *MRC Biostatistics Unit*.
- Wright, G.W., and Simon, R.M. (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics*, **19**, 18.

Diagnostic on controlled calibration model with replicates on both variables

Betsabé G. Blas Achic¹, Heleno Bolfarine¹ and Hugo Lachos²

¹ Departamento de Estatística, Universidade de São Paulo, São Paulo, Brasil

² Departamento de Estatística, Universidade Estadual de Campinas, São Paulo, Brasil

Abstract: In this work we generalize the controlled calibration model by assuming replication on both variables. Likelihood-based methodology is used to estimate the model parameters and the Fisher information matrix is used to construct a confidence interval for the unknown value of the regressor variable. We study the local influence diagnostic method for which it is developed the EM algorithm. A simulation study is carried out to assess the effect of the measurement error on the estimation of the parameter of interest. This new approach is illustrated with an example.

Keywords: local influence; regression model; linear calibration model; measurement error model; Berkson model.

1 Introduction

The calibration problem has two stages. In the first stage, the calibration experiment, one observes measurements (X_i, Y_{ij}) ($i = 1, \dots, n$, and $j = 1, \dots, m_i$) such that

$$Y_{ij} = \alpha + \beta X_i + \epsilon_{ij}, \quad i = 1, \dots, n, \text{ and } j = 1, \dots, m_i, \quad (1)$$

where ϵ_{ij} are independent and identically distributed (i.i.d.) $N(0, \sigma_\epsilon^2)$. In the second stage, the prediction stage, one observes Y_{0i} , $i = 1, \dots, k$ on a fixed unknown value X_0 such that

$$Y_{0i} = \alpha + \beta X_0 + \epsilon_{0i}, \quad i = 1, \dots, k. \quad (2)$$

We assume that ϵ_{ij} and ϵ_{0i} are i.i.d. $N(0, \sigma_\epsilon^2)$. The variable X_i takes fixed value. The model parameters are α, β, X_0 and σ_ϵ^2 , and the main interest is to estimate the value X_0 .

In this paper, the model (1) and (2) will be called *usual calibration model with replicates on both variable (U-M)*.

The maximum likelihood estimators of the U-M are given by

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}, \quad \hat{\beta} = \frac{S_{XY}}{S_{XX}}, \quad \hat{X}_0 = \frac{\bar{Y}_0 - \hat{\alpha}}{\hat{\beta}}, \quad (3)$$

$$\hat{\sigma}_\epsilon^2 = \frac{1}{k + \sum_{i=1}^n m_i} \left[\sum_{i=1}^n \sum_{j=1}^{m_i} (Y_{ij} - \hat{\alpha} - \hat{\beta} X_i)^2 + \sum_{i=1}^k (Y_{0i} - \bar{Y}_0)^2 \right], \quad (4)$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $\bar{Y} = \frac{1}{N} \sum_{i=1}^n \sum_{j=1}^{m_i} Y_{ij}$, $S_{XY} = \frac{1}{N} \sum_{i=1}^n \sum_{j=1}^{m_i} (X_i - \bar{X})(Y_{ij} - \bar{Y})$, $S_{XX} = \frac{1}{N} \sum_{i=1}^n m_i (X_i - \bar{X})^2$, $\bar{Y}_0 = \frac{1}{k} \sum_{i=1}^k Y_{0i}$, $N = \sum_{i=1}^n m_i$, and

$$v(\hat{X}_0) = \frac{\sigma_\epsilon^2}{\beta^2} \left[\frac{1}{k} + \frac{1}{N} + \frac{(\bar{X} - X_0)^2}{NS_{XX}} \right]. \quad (5)$$

In Blas, B.G. et al.(2007) it was proposed the so called *homoscedastic controlled calibration model*, where \hat{X}_i is a pre-fixed target value of a predictor variable (x_i), but the unobserved true value x_i can differ from the target value, it is to say

$$x_i = X_i - u_i, \quad i = 1, \dots, n, \quad (6)$$

here u_i is Berkson-type measurement error. We propose a generalization of the homoscedastic controlled calibration model by considering the usual model (1 and 2) and (6), lying in *the controlled calibration model with replicated error in the response variable (P-M)*.

This paper is organized in the following manner: In Section 2, it is presented the P-M, the parameter estimation methods and hypothesis testing. Section 3 presents results of simulation studies. In section 4, we provide a brief sketch of the local influence measures developed via Zhu, H. and Lee, S. (2001) approach. Four different perturbation schemes are considered. Section 5 presents illustrative examples with a real data set. In the next section some conclusion derived from this work is presented.

2 Controlled calibration model with replicated error-in-variable

In Blas, B.G. et al.(2007) is defined the homoscedastic controlled calibration model. This model is developed by considering controlled variables and assuming that the measurement errors have equal variances. In this work we generalize it by considering replication on both variables. Thus, we propose the following model:

$$Y_{ij} = \alpha + \beta X_i + (\epsilon_{ij} + \beta u_i) \quad i = 1, \dots, n, \text{ and } j = 1, \dots, m_i, \quad (7)$$

$$Y_{0i} = \alpha + \beta X_0 + \epsilon_{0i}, \quad i = 1, \dots, k. \quad (8)$$

with the following suppositions about the random errors: ϵ_{ij} , $\epsilon_{0i} \stackrel{ind}{\sim} N(0, \sigma_\epsilon^2)$; $u_i \stackrel{ind}{\sim} N(0, \sigma_u^2)$; $cov(u_i, \epsilon_{ij}) = 0$ for all i, j , and $cov(u_i, \epsilon_{0j}) = 0$ for all i, j . The model parameters are $\alpha, \beta, X_0, \sigma_\epsilon^2$ and σ_u^2 and the main interest is to estimate the quantity X_0 .

2.1 Maximum likelihood estimators

The maximum likelihood (ML) estimators for the homoscedastic controlled calibration model with replication on both variable (*P-M*) are given by:

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}, \quad \hat{X}_0 = \frac{\bar{Y}_0 - \hat{\alpha}}{\hat{\beta}}, \quad \hat{\beta} = \frac{\sum_{i=1}^n \sum_{j=1}^{m_i} (X_i - \bar{X})(Y_{ij} - \bar{Y})}{\sum_{i=1}^n m_i (X_i - \bar{X})^2} \quad (9)$$

$$\hat{\sigma}_\delta^2 = \frac{\sum_{i=1}^n \sum_{j=1}^{m_i} \{[(Y_{ij} - \bar{Y}) - \hat{\beta}(X_i - \bar{X})]^2 - \hat{\sigma}_\epsilon^2\}}{\sum_{i=1}^n m_i \hat{\beta}^2}, \quad \hat{\sigma}_\epsilon^2 = \frac{\sum_{i=1}^k (Y_{0i} - \bar{Y}_0)^2}{k}, \quad (10)$$

where

$$\bar{Y} = \frac{\sum_{i=1}^n \sum_{j=1}^{m_i} Y_{ij}}{\sum_{i=1}^n m_i}, \quad \bar{X} = \frac{\sum_{i=1}^n X_i}{n}, \quad Y_0 = \frac{\sum_{i=1}^k Y_{0i}}{k}. \quad (11)$$

We can observe that $\hat{\sigma}_\delta^2$ can supply negative values when

$$\frac{\sum_{i=1}^n \sum_{j=1}^{m_i} [(Y_{ij} - \bar{Y}) - \beta(X_i - \bar{X})]^2}{\sum_{i=1}^n m_i} < \frac{\sum_{i=1}^k (Y_{0i} - Y_0)^2}{k}. \quad (12)$$

The variance of \hat{X}_0 , derived from the Fisher information matrix, is given by

$$V(\hat{X}_0) = \frac{\sigma_\epsilon^2}{\beta^2} \left[\frac{1}{k} + \frac{\gamma}{N\sigma_\epsilon^2} + \frac{\gamma}{\sigma_\epsilon^2} \frac{(\bar{X} - X_0)^2}{\sum_{i=1}^n m_i (X_i - \bar{X})^2} \right], \quad \text{where } \gamma = \beta^2 \sigma_\delta^2 + \sigma_\epsilon^2. \quad (13)$$

2.2 Inference

Confidence interval: The approximated confidence interval for X_0 with a confidence level $(1 - \alpha)$, is given by

$$\left(\hat{x}_0 - z_{\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{X}_0)}, \hat{x}_0 + z_{\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{X}_0)} \right), \quad (14)$$

where $z_{\frac{\alpha}{2}}$ is the quantile of order $(1 - \frac{\alpha}{2})$ of the standard normal distribution.

Testing hypothesis: the hypotheses $H_0 : X_0 = X_{00}$ can be tested by using interval (14).

2.3 EM estimators

For the current value θ , the E-step of the EM-type algorithm requires the evaluation of $Q(\theta|\hat{\theta}) = E(l_c(\theta|\mathbf{Y}, \mathbf{Y}_0, \mathbf{x})|\mathbf{Y}, \mathbf{Y}_0, \hat{\theta})$.

Let \hat{x}_i and \hat{x}_i^2 , $i = 1, \dots, n$ given by

$$\hat{x}_i = X_i + \frac{\beta\sigma_u^2}{\gamma_i} \sum_{j=1}^{m_i} (Y_{ij} - \alpha - \beta X_i) \quad (15)$$

$$\hat{x}_i^2 = \frac{\sigma_\epsilon^2 \sigma_u^2}{\gamma} + X_i^2 + \frac{\beta^2 \sigma_u^4}{\gamma_i^2} \left(\sum_{j=1}^{m_i} (Y_{ij} - \alpha - \beta X_i) \right)^2 + \frac{2\beta\sigma_u^2 X_i}{\gamma_i} \sum_{j=1}^{m_i} (Y_{ij} - \alpha - \beta X_i), \quad (16)$$

where $\gamma_i = \sigma_\epsilon^2 + m_i \beta^2 \sigma_u^2$.

The M-step requires the maximization of $Q(\theta|\hat{\theta})$ with respect to θ . The closed-form equation for the M-step are given by

$$\begin{aligned} \hat{x}_0 &= \frac{Y_0 - \hat{\alpha}}{\hat{\beta}}, \quad \hat{\beta} = \frac{\sum_{i=1}^n \sum_{j=1}^{m_i} \hat{x}_i (Y_{ij} - \alpha)}{\sum_{i=1}^n m_i \hat{x}_i^2}, \quad \hat{\alpha} = \frac{\sum_{i=1}^n \sum_{j=1}^{m_i} Y_{ij} - \hat{\beta} \sum_{i=1}^n m_i \hat{x}_i}{\sum_{i=1}^n m_i} \\ \hat{\sigma}_u^2 &= \frac{1}{n} \sum_{i=1}^n (\hat{x}_i^2 + X_i^2 - 2\hat{x}_i X_i) \\ \hat{\sigma}_\epsilon^2 &= \frac{1}{k + \sum_{i=1}^n m_i} \left[\sum_{i=1}^n \sum_{j=1}^{m_i} \left((Y_{ij} - \alpha)^2 + \beta^2 \hat{x}_i^2 - 2\beta \hat{x}_i (Y_{ij} - \alpha) \right) + \sum_{i=1}^k (Y_{0i} - \alpha - \beta X_0)^2 \right] \end{aligned} \quad (17)$$

Substituting for \hat{x}_i and \hat{x}_i^2 from equations (15) and (16), respectively, into equation (17) we have that,

$$\hat{\sigma}_u^2 = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{\sigma_\epsilon^2 \sigma_u^2}{\gamma_i} + \frac{\beta^2 \sigma_u^2}{\gamma_i} \left(\sum_{j=1}^{m_i} (Y_{ij} - \alpha - \beta X_i) \right)^2 \right\}. \quad (18)$$

Thus, from (18) we easily see that the EM estimator of σ_u^2 (17) is always positive while the ML estimator of σ_u^2 (10) can assume both positive and negative values.

3 Simulation study

The aim of this section is to study the performance of the maximum likelihood estimators of the proposed model (P-M) and verify the impact by considering erratically the usual model.

It was considered 1000 samples generated from the new approach. In all samples, the value of the parameters α and β were 0.1 and 2, respectively. The range of values for the controlled variable was $[0,2]$. The fixed values for the controlled variable were $X_i = 2(i-1)/(n-1)$, $i = 1, \dots, n$, and the parameter values X_0 were 0.01 and 0.8. It was considered $\sigma_\epsilon^2 = 0.04$ and $\sigma_u^2 = 0.1$. On the first and second stages we consider the sample of sizes $n = 5, 20, 100$, $k = 2, 20$ and $m = 2, 10$.

The empirical mean bias is given by $\sum_{j=1}^{1000} (\hat{X}_0 - X_0)/1000$ and the empirical mean squared error (MSE) is given by $\sum_{j=1}^{1000} (\hat{X}_0 - X_0)^2/1000$. The mean estimated variance of \hat{X}_0 is given by $\sum_{j=1}^{1000} \hat{V}(\hat{X}_0)/1000$. The theoretical variances of \hat{X}_0 denoted as $V(\hat{X}_0)$ is referred to the expression (13) evaluated on the relevant parameter values.

Table 1 presents the empirical bias and MSE, the estimated variance of X_0 from the proposed and usual model, the theoretical variance $V(\hat{X}_0)$, the covering percentages % and the confidence interval amplitudes A constructed with a 95% confidence level for the parameter X_0 .

Figure 1 presents the percentage of rejection (%) of the testing hypothesis $H_0 : X_0 = X_{00}$, where $X_{00} \in [0, 2]$ and for each X_{00} it is used 1000 samples generated from the P-M with $\alpha = 0.1$, $\beta = 2$, $X_0 = 0.8$, $\sigma_\epsilon^2 = 0.04$, $n = 5$, $k = 2$, $m = 2$ and $\sigma_u^2 = 0.01$ (in Figure 1(a) and with $\sigma_u^2 = 0.1$ (in Figure 1(b)). The testing hypothesis is given according to Wald test for a level of significance 0.05.

We observe in Table 1 that for all X_0 , n , k and m the estimated variance of the proposed model is very close to the theoretical variance, whereas, the estimated variance from the usual model does not approach to the theoretical value. Also, analyzing the amplitude we observe that for all X_0 , n , k and m when it is adopted erratically the usual model the amplitudes present large values compared with the results from the proposed model. We observe

TABLE 1. Empirical bias and mean squared error, the mean estimated variance of \hat{X}_0 , theoretical variance, covering percentage (%) and amplitude (A) of the intervals with a 95% confidence level for the parameter X_0 , for $\sigma_u^2 = 0.1$.

X_0	(n, k, m)	Empirical		Mean of $V(\hat{X}_0)$		P-M $V(\hat{X}_0)$	U-M		P-M	
		Bias	MSE	U-M	P-M		%	A	%	A
0.01	(5, 2, 2)	-0.053	0.098	0.061	0.039	0.038	81.50	0.850	71.58	0.660
	(5, 2, 5)	-0.049	0.091	0.050	0.020	0.018	81.98	0.793	59.84	0.479
	(5, 20, 2)	-0.047	0.087	0.016	0.034	0.033	55.64	0.424	69.66	0.596
	(5, 20, 5)	-0.049	0.089	0.011	0.014	0.013	48.06	0.355	52.38	0.395
	(20, 2, 2)	-0.015	0.027	0.060	0.016	0.015	99.40	0.939	83.60	0.469
	(20, 2, 5)	-0.015	0.026	0.055	0.009	0.009	99.28	0.900	70.66	0.352
	(20, 20, 2)	-0.015	0.022	0.011	0.011	0.011	82.78	0.398	82.24	0.396
	(20, 20, 5)	-0.015	0.022	0.008	0.005	0.004	75.98	0.341	63.50	0.260
	(100, 2, 2)	-0.002	0.009	0.056	0.007	0.007	100.00	0.923	82.58	0.305
	(100, 2, 5)	-0.003	0.009	0.055	0.006	0.006	100.00	0.916	73.90	0.262
	(100, 20, 2)	-0.004	0.005	0.007	0.003	0.003	98.24	0.326	86.02	0.201
	(100, 20, 5)	-0.002	0.005	0.006	0.001	0.001	97.48	0.304	71.76	0.144
0.80	(5, 2, 2)	-0.005	0.033	0.043	0.016	0.017	91.30	0.734	75.28	0.443
	(5, 2, 5)	-0.009	0.030	0.041	0.010	0.010	92.56	0.723	65.48	0.347
	(5, 20, 2)	-0.009	0.030	0.006	0.011	0.012	59.42	0.274	70.14	0.355
	(5, 20, 5)	-0.008	0.026	0.005	0.004	0.005	56.44	0.249	54.34	0.237
	(20, 2, 2)	-0.001	0.011	0.052	0.008	0.008	99.88	0.876	82.84	0.326
	(20, 2, 5)	-0.005	0.011	0.052	0.006	0.006	99.98	0.881	73.28	0.277
	(20, 20, 2)	-0.003	0.006	0.006	0.004	0.004	91.94	0.289	83.26	0.226
	(20, 20, 5)	-0.002	0.006	0.005	0.002	0.002	91.42	0.282	67.88	0.159
	(100, 2, 2)	-0.001	0.006	0.055	0.006	0.006	100.00	0.913	78.42	0.257
	(100, 2, 5)	-0.002	0.006	0.055	0.005	0.005	100.00	0.913	72.66	0.237
	(100, 20, 2)	-0.001	0.002	0.005	0.001	0.001	99.96	0.291	88.30	0.130
	(100, 20, 5)	0.000	0.002	0.005	0.001	0.001	99.94	0.290	80.26	0.106

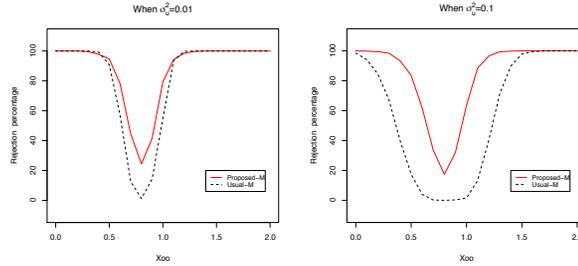


FIGURE 1. Variation of the percentage of rejection (%) of the testing hypothesis $H_0 : X_0 = X_{00}$ for $\sigma_u^2 = 0.01$ (Figure 1(a)) and $\sigma_u^2 = 0.1$ (Figure 1(b)).

that when m increases the amplitude decrease. Also, in most cases, for all n, m and X_0 when it is adopted erratically the U-M, the amplitudes decrease very much as the size of k increases. This causes similar behavior for the covering percentage.

In Figure 1(a) the percent of rejection from P-M and U-M around the null hypothesis values $X_0 = 0.8$ reveals an overall increase in the percent of rejection. We also observe that the percent of rejection from P-M is much more than one from U-M. Moreover, when there is an increasing in the size of the variance σ_u^2 (Figure 1(b)) it has much greater impact on the percent of rejection from both model.

4 Local influence of the proposed model

In this work, we use the methodology developed by Zhu, H. and lee, S.(2001), an unified method for conducting local influence analysis for general statistical models with missing data based on the Q-function used within the EM implementation.

As defined in Zhu, H. and lee, S.(2001) the assessment of influential cases is based on the visual inspection of the $\{M(0)_l, l = 1, \dots, g\}$ plotted against the index l . We consider four different perturbation schemes for the P-M. The case weight perturbation, response variable perturbation, perturbation of the variance σ_ϵ^2 , perturbation of the variance σ_u^2 .

5 Application

In this section we present some applications by using the data supplied by the chemical laboratory of the "Instituto de Pesquisas Tecnológicas (IPT)" - Brazil. We also consider the U-M in order to observe the performance of the P-M.

Table 2 (a) presents the fixed values of concentration of the standard solutions and the corresponding intensities for the bario element, which are supplied by the plasma spectrometry method. This data is referred to as the first stage of the calibration model. Table 2 (b) presents the intensities corresponding to 3 sample solutions from the sample A and B. This data is referred to as the second stage of the calibration model.

Figure 2 displays plot of $M(0)_j, j = 1, \dots, 15$ related to the samples A and B.

All estimates are computed considering the intensities divided by 10000 on the data given in Table 2 .

Table 3 (a) and (b) describe the ML and EM estimates of $\alpha, \beta, X_0, V(\hat{X}_0), \sigma_\delta^2$ and the confidence interval amplitude $U(X_0)$ from the P-M of the samples A and B for the bario chemical element for the complete data and when it is regarded the influential data detected in the Figure 2, respectively. It is also presented the estimates from the U-M. The estimates of the variance of \hat{X}_0 are computed using the relevant expressions (13) and (5). The amplitude $U(X_0)$ is given by the product of the squared root of the estimated variance of \hat{X}_0 and 1.96.

In Tables 3 (a) and (b) we can observe that the ML and EM estimates of α and β supplied by the U-M is equal to the P-M, this occurs because the expression of the estimators $\hat{\alpha}$ and $\hat{\beta}$ of both models are equal and they only depend on the first stage of the calibration model. The ML estimates are different of the EM estimates. We can observe that the estimate of the concentration of the sample B is outside of the variation range of the standard solution concentrations. The estimates computed on data without the influence data, described in Table 3 (b), are different compared with

the estimates related to the complete data. In Table 3 (b) we verify that the estimate of the variance of \hat{X}_0 and the amplitude $U(X_0)$ from the U-M are greater than the estimates supplied by the P-M.

In the Figures 2 (a) and (b) we observe that the intensity 1926319.8, corresponding to the concentration 1.06 from the first step is detected as influential data in the weight perturbation and perturbation of the variance σ_ϵ^2 schemes. In Table 3 (b) without considering this influential data, we observe that the estimates of α , β , and X_0 (in both samples) are slightly different from the results obtained by considering all data, but in the sample A the estimates of $V(\hat{X}_0)$ and $U(X_0)$ differs from the results obtained by considering all data. We note that in the sample B the estimate of σ_u^2 is negative then we can not compute $V(\hat{X}_0)$ and $U(X_0)$, in this case we only have EM estimates.

Testing the hypothesis $H_0 : X_0 = 0.2$ and $H_0 : X_0 = 0.149$ for the sample A, considering the significance level 5% and using the EM estimates lead to $W = 474.9612$ ($p = 0$) and $W = 0.01663$ ($p = 0.8974$), respectively. On the other hand, testing the hypothesis $H_0 : X_0 = 1.6$ and $H_0 : X_0 = 1.48$ for the sample B, by regarding the significance level 5% and the EM estimates we have that $W = 562.7771$ ($p = 0$) and $W = 0.6692$ ($p = 0.4133$). We observe that when the null hypothesis are the nearest number to the EM estimate the P-values are more than the significance level, then the result is statistically non-significant. When the null hypothesis are the round up ML estimates the results are statistically significant.

TABLE 2. (a): concentration (mg/g) and intensity of the standard solutions of bario element.(b): intensity of the sample solutions A and B of bario element.

Intensity	(a)				(b)	
	$X_1 = 0.1$	$X_2 = 0.2$	$X_3 = 0.5$	$X_4 = 1.06$	Sample A	Sample B
	185082.2	373543.2	913829.3	1895804.6	279034.2	2640425.3
	184583.0	375166.9	894229.6	1926319.8	279562.1	2661015.0
	184906.3	369481.1	911759.2	1886632.8	278462.2	2639452.0

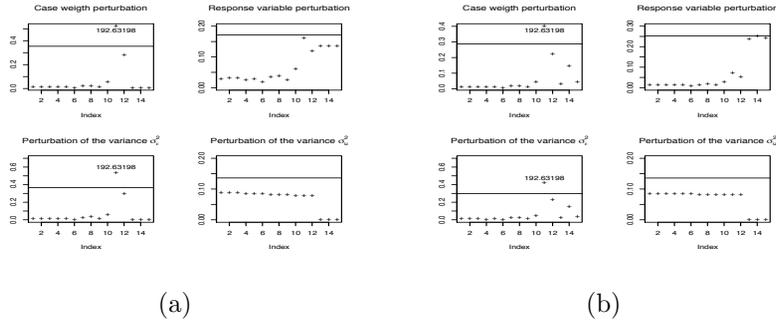


FIGURE 2. Index plot of $M(0)_j$ for four types of perturbation. Figure 2 (a) related to the sample A and Figure 2 (b) related to the sample B.

TABLE 3. Estimates of α , β , X_0 , $V(\hat{X}_0)$ and the confidence interval amplitude $U(X_0)$ from the usual and proposed model for the samples A and B of chromo element. (a) computed with complete data and (b) computed without the influential data.

		(a)			
Sample	Parameter	Method			
		U-M		P-M	
		ML	EM	ML	EM
A	α	1.139	1.139	1.139	1.139
	β	1.786E+02	1.786E+02	1.786E+02	1.786E+02
	X_0	1.499E-01	1.499E-01	1.499E-01	1.499E-01
	σ_y^2	-	-	3.298E-05	5.749E-07
	$V(\hat{X}_0)$	1.260E-05	1.260E-05	4.747E-06	1.271E-05
	$U(X_0)$	6.957E-03	6.957E-03	4.270E-03	6.987E-03
B	α	1.139	1.139	1.139	1.139
	β	1.786E+02	1.786E+02	1.786E+02	1.786E+02
	X_0	1.476	1.476	1.476	1.476
	σ_y^2	-	-	-1.343E-05	4.537E-07
	$V(\hat{X}_0)$	3.352E-05	3.352E-05	-	3.394E-05
	$U(X_0)$	1.135E-02	1.135E-02	-	1.142E-02

		(b)			
Sample	Parameter	Method			
		U-M		P-M	
		ML	EM	ML	EM
A	α	-6.197	1.413	-6.197	1.413
	β	173.152	1.774E+02	173.152	1.774E+02
	X_0	0.197	1.493E-01	0.197	1.493E-01
	σ_y^2	-	-	0.003	4.970E-06
	$V(\hat{X}_0)$	0.001	6.596E-06	0.0004	5.412E-06
	$U(X_0)$	0.066	5.034E-03	0.041	4.560E-03
B	α	-6.197	1.413	-6.197	1.413
	β	173.152	1.774E+02	173.152	1.774E+02
	X_0	1.564	1.484	1.564	1.484
	σ_y^2	-	-	0.003	2.006E-06
	$V(\hat{X}_0)$	0.003	2.405E-05	0.003	2.391E-05
	$U(X_0)$	0.111	9.612E-03	0.108	9.584E-03

6 Concluding remarks

Simulation study reveals that when it is considered the error variance σ_y^2 , the mean estimated variance of \hat{X}_0 obtained using the U-M moves away from the theoretical value. U-M shows a far greater storage confidence interval amplitude than the P-M. This emphasizes the storage uncertainty related to X_0 concentration. In general, the proposed model is sensible to the presence of the independent measurement error and gives better results in contrast to the U-M results. In the example above, we observe that one observation is detected as influential data, and the estimation without it differs from the results considering all data. It is observed that despite the fact that the U-M, does not consider measurement errors, the amplitude is greater than the one obtained by the new approach.

Acknowledgements

Betsabé G. Blas Achic has been supported by CNPq .

References

- Blas, B.G; Sandoval, M.C.; Satomi, O.Y. (2007). Homoscedastic controlled calibration model. *Technometrics*, **21**, 145.
- Zhu, H. and lee, S. (2001). Local influence for incomplete-data models. *Journal of the Royal Statistical Society, Series B*, **63**, 111.

Mapping brain activity through spatiotemporal smoothing

A.W. Bowman¹, C. Ferguson¹, N. Dean¹, and J. Gross²

¹ Department of Statistics, University of Glasgow, Glasgow G12 8QQ, UK

² Department of Psychology, University of Glasgow, Glasgow G12 8QQ, UK ,
E-mail: adrian@stats.gla.ac.uk

Abstract: Magnetoencephalography involves the recording of electrical activity in the brain over time at a large number of spatial locations on a helmet placed over a human skull. This results in very large datasets with both spatial and temporal structure. Spatiotemporal smoothing is employed to identify and characterise stimulated events in brain activity. Technical issues involve the determination of distance between locations on the surface of the helmet and the efficient implementation of simultaneous smoothing across space and time. The resulting reduction in noise gives clear indications of spatial and temporal patterns which can provide the basis of further analysis.

Keywords: spatiotemporal; brain activity; smoothing

1 Introduction

Magnetoencephalography (MEG) is an imaging technique used to measure the electrical activity in the brain through sensitive recording devices embedded in a helmet placed over a human skull. Hämäläinen *et al.* (1993) give the technical background, while a representation of a helmet with 246 sensors is shown in Figure 1.

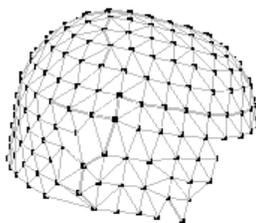


FIGURE 1. The helmet with 246 sensors.

The sensors record electrical activity at very high time resolution and subjects are given a repeated stimulus. This results in very large datasets with both spatial and temporal structure. The top left panel of Figure 2 shows one particular replicate of 114 collected on a single subject. Considerable

high frequency patterns are evident, while close inspection of individual sensors indicates low frequency signals at around 10Hz.

A common approach to analysis is based on the mean signal across replicates, at each time point and for each sensor. The resulting patterns are displayed in the middle left panel of Figure 2, where reduction in variance and the presence of a brain response to the stimulus, located at 0 on the time axis, are both now apparent. However, the expectation that the mean response will vary smoothly over both time and space suggests that the application of appropriate spatiotemporal smoothing techniques has the potential to identify and characterise the patterns of brain activity more effectively.

2 Methods

In this spatial domain, standard forms of smoothing are not appropriate because smoothing must take place across the surface of the helmet. Distance between sensor locations cannot be measured in a Euclidean manner but by the length of a geodesic curve along the surface of the helmet. This distance can be calculated by an appropriate reorientation of the co-ordinate axes and the identification of the sensor locations which lie close to this geodesic curve. One dimensional smoothing along an axis corresponding to the geodesic direction, using the projections of the adjacent sensor locations onto this axis, provides a simple but effective estimate of the geodesic curve. The length of this curve can be easily measured to compute distance. Spatial smoothing on the plane can be implemented by a variety of straightforward techniques. Spatial smoothing on the helmet surface can be implemented by appropriate use of the distance between the sensor locations. A very simple approach constructs an estimate of the mean response at sensor s , from the n sensor data $y_i, i = 1, \dots, n$, as

$$\frac{\sum_{i=1}^n w_{si} y_i}{\sum_{i=1}^n w_{si}}, \quad w_{si} = \exp\{-0.5d_{si}^2/h^2\},$$

where d_{si} denotes the geodesic distance between sensors s and i . This simple local mean can easily be replaced by a local linear estimator, with superior edge properties. Fan & Gijbels (1996) give the details in the setting of planar covariates.

Temporal smoothing at each sensor is simple to implement by any convenient smoothing procedure. However, simultaneous smoothing over space and time involves a three-dimensional covariate space which is more unusual. Bowman *et al.* (2009) discuss this in the context of spatiotemporal environmental data and show that smoothing across space, followed by smoothing the spatial fitted values across time at each spatial location, provides a very efficient procedure which enjoys the same asymptotic properties as direct three-dimensional smoothing. The array structure of the

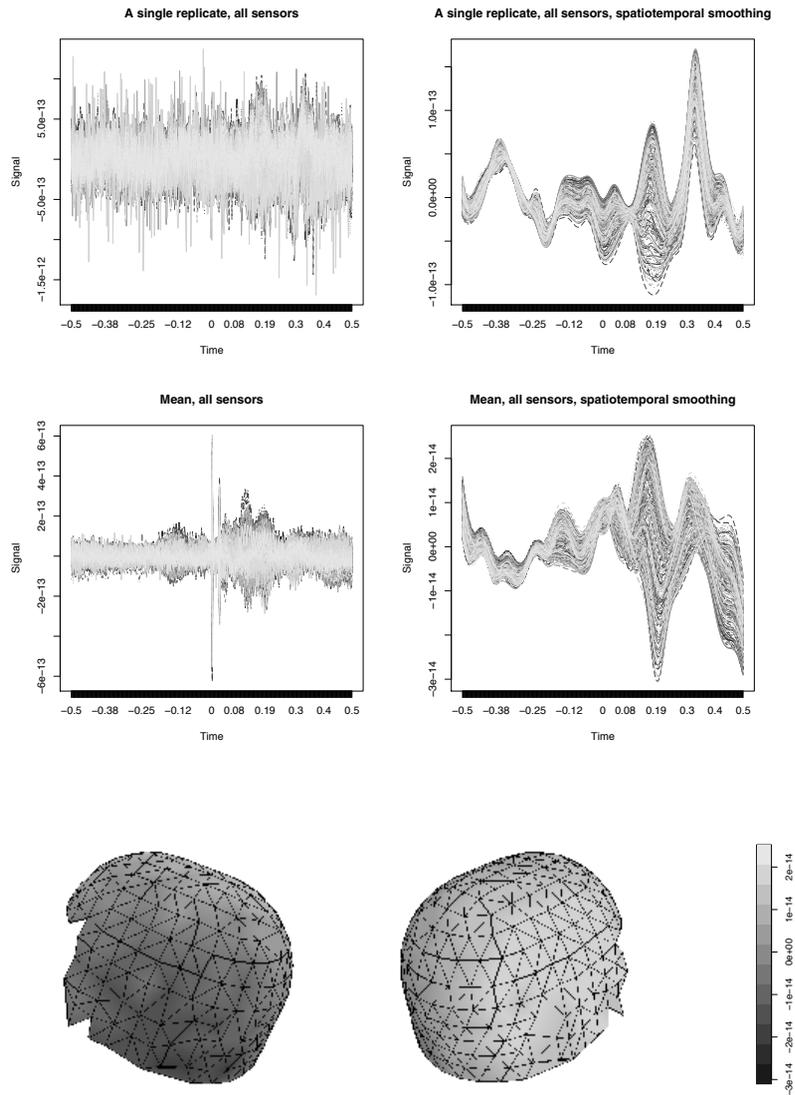


FIGURE 2. The top panels show the signals over time for all sensors, for a single particular replicate, with the raw data on the left and the smoothed version on the right. The middle panels show the mean of all replicates over time for all sensors, with the raw data on the left and the smoothed version on the right. The lower panels display the smoothed mean signal as a colour coded surface over the helmet, from left and right views, at a particular time point.

data allows the estimate to be expressed as $S_s Y S_t^T$, where Y denotes the data in matrix form with rows for location and columns for time and S_s, S_t denote spatial and temporal smoothing matrices. This places the problem in the GLAM framework of Currie *et al.* (2006).

3 Results

The right hand panels of Figure 2 show the results of spatiotemporal smoothing on the signal traces. The high frequency patterns are effectively suppressed leaving the low frequency signal, at approximately 10Hz, more apparent. In particular, the post-stimulus response is identified clearly at both individual replicate and mean levels.

The spatial patterns are more effectively represented by colour coding the surface of the helmet with the mean response at a particular time point. An animation across time, using a slider control, provides a very effective characterisation of the principal locations of the post-stimulus response as well as its timing. The lower panels of Figure 2 illustrate this in a static plot at a particular time point in the middle of the post-stimulus response pattern, using two orientations to show the full helmet surface.

While a common approach is to analyze the mean signal across replicates, it is also of interest to investigate the post-stimulus response for individual replicates. There is a large amount of variation evident across replicates which makes it difficult in some cases to identify where post-stimulus ‘events’ occur. Where the locations of events can be clearly identified, there is also variation in the timing of these. It is therefore necessary to examine possible criteria that will allow the occurrence and timing of events to be identified. A simple approach is to characterise this by an increase in the mean and/or spatial variance of the post-stimulus response. This is expected to occur between 5 and 200 ms. However, the timing is likely to be different for each replicate.

Variation in individual replicates can simply be a result of eye movement, independent of the stimulus, while trend in the response signal can also be introduced by external background noise. The smooth estimates for each replicate were therefore de-trended using linear regression. The top four panels of Figure 3 display de-trended smoothed responses for four individual replicates, with lines to identify the time window of interest for events. Large increases in the mean and spatial variance are evident in the first two plots within this period. However there is little evidence of any event in the second two replicates. Threshold values were therefore used to identify the presence and timing of events. A simple criterion is to compare the maximum values of the mean and spatial variance, in the time window of interest, to the corresponding maximum values of these summary measures in the baseline, pre-stimulus time period. If the largest post-stimulus mean exceeds this threshold value, the time of the event is

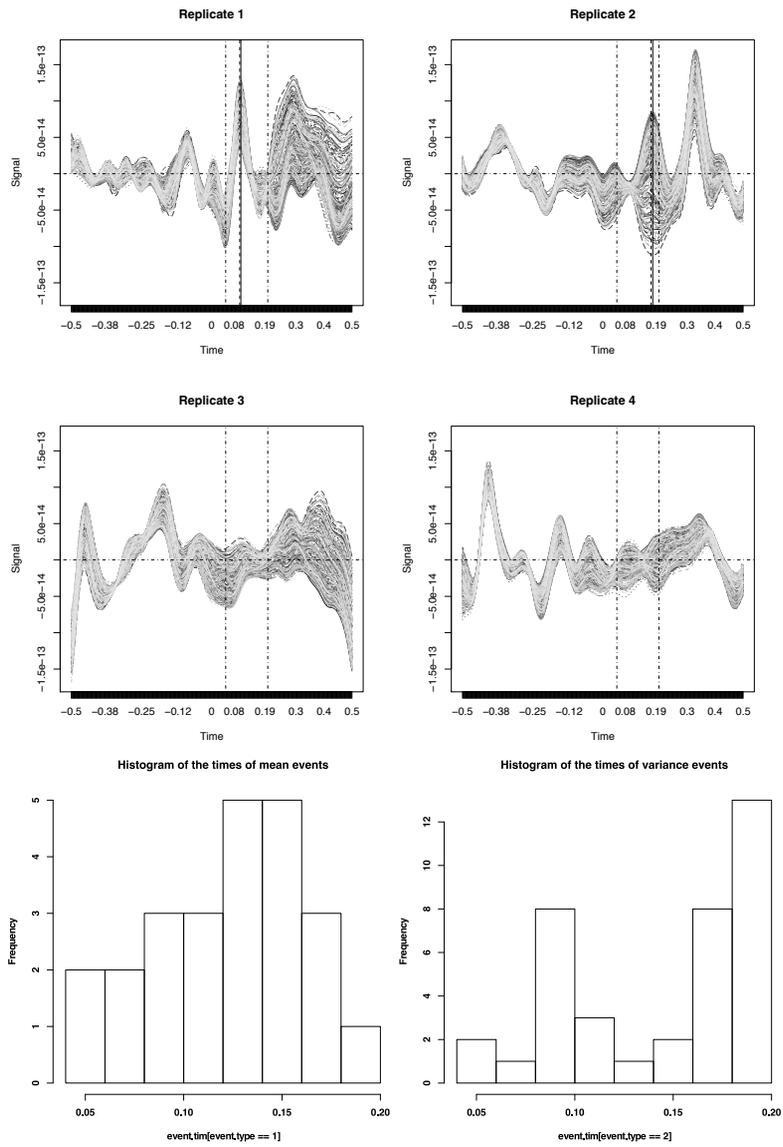


FIGURE 3. The top four panels show de-trended smoothed responses for four individual replicates, with vertical (dot-dashed) lines to identify the time window of interest for events. In the top two panels, events are identified by the mean (full lines) and the variance (dashed lines). The lower two panels show histograms of the times of events identified by the mean (left) and the variance (right).

noted, as highlighted by the full line in the top two panels of Figure 3. If the maximum mean value does not exceed the threshold then the maximum variance is considered relative to its pre-stimulus threshold and a similar process applied. The results for the variance are displayed as dashed lines on the top two panels of Figure 3. If neither threshold is exceeded then no event is recorded, as was the case in the middle two panels of Figure 3. Applying these criteria to all of the replicates identified events in 62 out of the 114 cases with most of these identified using the variance. The timings of these events are displayed for both summary measures in the bottom two panels of Figure 3 which indicate an approximately normal distribution for the mean, with most events occurring towards the middle of the time period, and a quite different and possibly bimodal distribution for the variance.

4 Discussion

The results show that spatiotemporal smoothing is effective in identifying and characterising the post-stimulus response in brain activity. This provides a very informative representation of individual replicates which can then provide the basis of further analysis, allowing for variation in the location, timing and strength of responses across replicates. The simple threshold criteria implemented here highlight the substantial variation between replicates in both the identification and the timing of events. Such criteria have been successful in identifying events and their timing within many of the replicates. However, further analysis is required to detect events in signals with less marked features and to compare the timings and sizes of events in individual replicates to the mean response signal.

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* (to appear).
- 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.
- Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman and Hall, London.
- Hämäläinen, M., Hari, R., Ilmoniemi, R., Knuutila, J. and Lounasmaa, O. V. (1993). Magnetoencephalography theory, instrumentation, and applications to noninvasive studies of signal processing in the human brain. In *Reviews of Modern Physics* **65**, 413–497.

Least-squares quadratic estimation in uncertain observation systems with different uncertainty probabilities

R. Caballero-Águila¹, A. Hermoso-Carazo²
and J. Linares-Pérez²

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, the state least-squares quadratic estimation problem from uncertain observations coming from multiple sensors is addressed. It is assumed that, at each sensor, the state is measured in the presence of additive white noise and that the uncertainty about the state being present or missing is characterized by a set of Bernoulli random variables whose probabilities are not necessarily the same for all the sensors. By defining suitable augmented state and observation vectors, the original quadratic estimation problem is reduced to the linear estimation problem of the augmented state.

Keywords: Least-squares estimation; Uncertain observations; Multiple sensors.

1 Introduction

There is a large class of real-world problems where the state appears in the observation in a random manner, such as problems where there are intermittent failures in the observation mechanism or fading phenomena in propagation channels. These situations are characterized by including in the observation equation, besides an additive noise, a multiplicative noise indicating the fact that the state is not always present in the observations but there is a nonzero probability that the measurement contains only noise (*uncertain observations*). Recently, the state least-squares linear estimation problem in systems with uncertain observations transmitted by multiple sensors whose statistical properties are assumed not to be the same for all the sensors has been studied by several authors under different approaches and hypotheses on the processes (see e.g. Hounkpevi and Yaz (2007), Jiménez-López et al. (2008) and references therein). In this paper an algorithm for the least-squares quadratic filter is proposed. To approach the quadratic estimation problem we use the technique proposed by De Santis et al. (1995), which consists of augmenting the signal and observation vectors, by aggregating their second-order powers defined by the

Kronecker product, thus obtaining a new augmented system and reducing the quadratic estimation problem from the original system to the linear estimation problem from the augmented system.

2 Hypotheses on the system

Consider the n -dimensional state process and the observation model:

$$\begin{aligned} x_k &= F_{k-1}x_{k-1} + w_{k-1}, & k \geq 1 \\ y_k^i &= \gamma_k^i H_k^i x_k + v_k^i, & k \geq 1, \quad i = 1, \dots, m \end{aligned} \quad (1)$$

where $\{w_k; k \geq 0\}$ is a white noise, $\{y_k^i; k \geq 1\}$ denotes the scalar observation process from the i -th sensor which, for $i = 1, \dots, m$, is perturbed by $\{v_k^i; k \geq 1\}$, a white process, and by $\{\gamma_k^i; k \geq 1\}$, a sequence of independent Bernoulli random variables. When $\gamma_k^i = 1$, the i -th sensor is assumed to work properly at time k , whereas $\gamma_k^i = 0$ means that the i -th sensor fails at time k . By denoting $y_k = (y_k^1, \dots, y_k^m)^T$, $H_k = (H_k^{1T}, \dots, H_k^{mT})^T$, $\Upsilon_k = \text{Diag}(\gamma_k^1, \dots, \gamma_k^m)$ and $v_k = (v_k^1, \dots, v_k^m)^T$, the observation model in (1) is rewritten in a compact form as follows:

$$y_k = \Upsilon_k H_k x_k + v_k, \quad k \geq 1. \quad (2)$$

In this paper, the least-squares (LS) quadratic (or second-order polynomial) estimators of the state x_k based on the observations till the instant k are obtained. For this purpose, we consider the random vectors $y_i^{[2]} = y_i \otimes y_i$ (\otimes denotes the Kronecker product) and, if $E[y_i^{[2]T} y_i^{[2]}] < \infty$, the required estimator is the orthogonal projection of x_k on the space of n -dimensional linear transformations of y_1, \dots, y_k and their second-order powers $y_1^{[2]}, \dots, y_k^{[2]}$. To assure the existence of the second-order moments of the vectors $y_i^{[2]}$, the following hypotheses are assumed:

(H1) The initial state x_0 is a random vector with known $\mu_0 = E[x_0]$, $\Sigma_0 = \text{Cov}[x_0]$, $\Sigma_0^{(3)} = \text{Cov}[x_0, x_0^{[2]}]$ and $\Sigma_0^{(4)} = \text{Cov}[x_0^{[2]}]$.

(H2) The state noise $\{w_k; k \geq 0\}$ is a zero-mean white sequence with known $Q_k = \text{Cov}[w_k]$, $Q_k^{(3)} = \text{Cov}[w_k, w_k^{[2]}]$ and $Q_k^{(4)} = \text{Cov}[w_k^{[2]}]$, $\forall k \geq 0$.

(H3) For $i = 1, \dots, m$, the noise $\{\gamma_k^i; k \geq 1\}$ is a sequence of independent Bernoulli variables with known probabilities, $P[\gamma_k^i = 1] = p_k^i$, $\forall k \geq 1$.

(H4) For $i = 1, \dots, m$, the sensor additive noises, $\{v_k^i; k \geq 1\}$, are zero-mean white processes and their moments, up to the 4th one, are known; we will denote $R_k = \text{Cov}[v_k]$, $R_k^{(3)} = \text{Cov}[v_k, v_k^{[2]}]$ and $R_k^{(4)} = \text{Cov}[v_k^{[2]}]$.

(H5) The initial state x_0 and the noise processes, $\{w_k; k \geq 0\}$, $\{\gamma_k^i; k \geq 1\}$ and $\{v_k^i; k \geq 1\}$, for $i = 1, \dots, m$, are mutually independent.

3 Least-squares quadratic estimation problem

In this section, the quadratic filter $\hat{x}_{k/k}^q$ of the state, x_k , from the observations (2) is obtained; for this, the following augmented state and observation vectors are defined by aggregating the original vectors with their second-order Kronecker powers: $\mathcal{X}_k = \begin{pmatrix} x_k^T & x_k^{[2]T} \end{pmatrix}^T$, $\mathcal{Y}_k = \begin{pmatrix} y_k^T & y_k^{[2]T} \end{pmatrix}^T$. Clearly, the space of n -dimensional linear transformations of $\mathcal{Y}_1, \dots, \mathcal{Y}_k$ is equal to the space of n -dimensional linear transformations of y_1, \dots, y_k and $y_1^{[2]}, \dots, y_k^{[2]}$. Then, the LS quadratic estimator, $\hat{x}_{k/k}^q$, is the LS linear estimator of x_k based on $\mathcal{Y}_1, \dots, \mathcal{Y}_k$. To obtain this linear estimator, firstly the relation between the augmented vectors \mathcal{X}_k and \mathcal{Y}_k is studied and the statistical properties of the augmented vectors are analyzed.

Augmented system. Using the Kronecker product properties and the system hypotheses it is deduced that the centered augmented vectors $X_k = \mathcal{X}_k - E[\mathcal{X}_k]$ and $Y_k = \mathcal{Y}_k - E[\mathcal{Y}_k]$ satisfy the following equations:

$$\begin{aligned} X_k &= \mathcal{F}_{k-1} X_{k-1} + W_{k-1}, \quad k \geq 1 \\ Y_k &= D_k^\gamma \mathcal{H}_k X_k + V_k, \quad k \geq 1 \end{aligned} \quad (3)$$

where $\mathcal{F}_k = \text{Diag}(F_k, F_k^{[2]})$, $\mathcal{D}_k^\gamma = \text{Diag}(\Upsilon_k, \Upsilon_k^{[2]})$, $\mathcal{H}_k = \text{Diag}(H_k, H_k^{[2]})$

$$W_k = \begin{pmatrix} w_k \\ (I + K) ((F_k x_k) \otimes w_k) + w_k^{[2]} - \text{vec}(Q_k) \end{pmatrix}$$

$$V_k = \begin{pmatrix} v_k \\ (I + K) ((\Upsilon_k H_k x_k) \otimes v_k) + v_k^{[2]} - \text{vec}(R_k) \end{pmatrix} + (D_k^\gamma - E[D_k^\gamma]) \mathcal{H}_k E[\mathcal{X}_k]$$

(I and K denote the identity and commutation matrices of appropriate dimensions and vec is the operator that vectorizes a matrix).

Properties of augmented vectors. In the following propositions the statistical properties of the processes involved in equation (3) are established.

Proposition 1. The noise $\{W_k; k \geq 0\}$ is a sequence of zero-mean, mutually uncorrelated random vectors with

$$E[W_k W_k^T] = \bar{Q}_k = \begin{pmatrix} Q_k & Q_k^{12} \\ Q_k^{12T} & Q_k^{22} \end{pmatrix}$$

where

$$Q_k^{12} = ((F_k \mu_k)^T \otimes Q_k) (I + K) + Q_k^{(3)} \text{ with } \mu_k = E[x_k] = F_{k-1} \mu_{k-1}, \quad k \geq 1$$

$$Q_k^{22} = (I + K) ((F_k s_k F_k^T) \otimes Q_k) (I + K) + ((F_k \mu_k)^T \otimes Q_k^{(3)T}) (I + K)$$

$$+ (I + K) ((F_k \mu_k) \otimes Q_k^{(3)}) + Q_k^{(4)}$$

$$\text{with } s_k = E[x_k x_k^T] = F_{k-1} s_{k-1} F_{k-1}^T + Q_{k-1}, \quad k \geq 1; \quad s_0 = \Sigma_0 + \mu_0 \mu_0^T.$$

Proposition 2. The random matrices $\{D_k^\gamma; k \geq 1\}$ are independent, $E[D_k^\gamma] = D_k^p = \text{Diag}(\Upsilon_k^p, \Upsilon_k^{[2]p})$, and they satisfy $E[D_k^\gamma \mathcal{H}_k X_k^T V_s^T] = 0, \forall s, k$.

Proposition 3. The noise $\{V_k; k \geq 1\}$ is a sequence of zero-mean, mutually uncorrelated random vectors with

$$E[V_k V_k^T] = \bar{R}_k = \begin{pmatrix} R_k & R_k^{12} \\ R_k^{12T} & R_k^{22} \end{pmatrix} + Cov[C_k^\gamma] \circ (\mathcal{H}_k E[\mathcal{X}_k] E[\mathcal{X}_k]^T \mathcal{H}_k^T)$$

where $E[\mathcal{X}_k] = (\mu_k^T, (vec(s_k))^T)^T$, $R_k^{12} = ((\Upsilon_k^p H_k \mu_k)^T \otimes R_k)(I + K) + R_k^{(3)}$,
 $R_k^{22} = (I + K) \left((E[C_{\Upsilon_k} C_{\Upsilon_k}^T] \circ (H_k s_k H_k^T)) \otimes R_k \right) (I + K) + R_k^{(4)}$
 $+ \left((\Upsilon_k^p H_k \mu_k)^T \otimes R_k^{(3)T} \right) (I + K) + (I + K) \left((\Upsilon_k^p H_k \mu_k) \otimes R_k^{(3)} \right)$,

\circ denotes the Hadamard product and $C_k^\gamma = \left(C_{\Upsilon_k}^T, C_{\Upsilon_k}^{[2]T} \right)^T$ with $C_{\Upsilon_k} = (\gamma_k^1, \dots, \gamma_k^m)^T$. Moreover $\{V_k; k \geq 1\}$ and $\{W_k; k \geq 0\}$ are uncorrelated.

In view of these properties, the linear estimators $\hat{X}_{k/k}$, of the state X_k based on the observations Y_1, \dots, Y_k , from which the quadratic estimators $\hat{x}_{k/k}^q$ are obtained, are calculated by applying the following recursive algorithm.

Filtering algorithm. The linear filter, $\hat{X}_{k/k}$, of X_k is given by

$$\hat{X}_{k/k} = \mathcal{F}_{k-1} \hat{X}_{k-1/k-1} + \Sigma_{k/k-1} \mathcal{H}_k^T D_k^p \Pi_k^{-1} [Y_k - D_k^p \mathcal{H}_k \mathcal{F}_{k-1} \hat{X}_{k-1/k-1}], \quad k \geq 1$$

with initial condition $\hat{X}_{0/0} = 0$, where the matrix Π_k satisfies

$$\Pi_k = (Cov[C_k^\gamma]) \circ (\mathcal{H}_k S_k \mathcal{H}_k^T) + D_k^p \mathcal{H}_k \Sigma_{k/k-1} \mathcal{H}_k^T D_k^p + \bar{R}_k, \quad k \geq 1$$

and the matrices S_k and $\Sigma_{k/k-1}$ are recursively calculated from

$$\begin{aligned} S_k &= \mathcal{F}_{k-1} S_{k-1} \mathcal{F}_{k-1}^T + \bar{Q}_{k-1}, \quad k \geq 1; \quad S_0 = E[X_0 X_0^T] \\ \Sigma_{k/k-1} &= \mathcal{F}_{k-1} \Sigma_{k-1/k-1} \mathcal{F}_{k-1}^T + \bar{Q}_{k-1}, \quad k \geq 1 \\ \Sigma_{k/k} &= \Sigma_{k/k-1} - \Sigma_{k/k-1} \mathcal{H}_k^T D_k^p \Pi_k^{-1} D_k^p \mathcal{H}_k \Sigma_{k/k-1}, \quad k \geq 1; \quad \Sigma_{0/0} = S_0. \end{aligned}$$

4 Numerical simulation example

To show the effectiveness of the proposed estimators, we ran a program in MATLAB, simulating at each iteration the state and the observed values and providing the linear and quadratic filtering estimates, as well as the corresponding error covariance matrices.

Consider a first-order autoregressive model,

$$x_k = 0.95x_{k-1} + w_{k-1}, \quad k \geq 1$$

where the initial state is a zero-mean Gaussian variable with $Var[x_0] = 1$ and $\{w_k; k \geq 0\}$ is a zero-mean white Gaussian noise with $Var[w_k] = 0.1$. Consider two sensors whose uncertain measurements,

$$y_k^i = \gamma_k^i x_k + v_k^i, \quad k \geq 1, \quad i = 1, 2$$

are perturbed by multiplicative noises $\{\gamma_k^i; k \geq 1\}$, $i = 1, 2$, which are sequences of independent Bernoulli variables with constant probabilities, $P[\gamma_k^i = 1] = p^i$, for all $k \geq 1$, and by independent additive zero-mean white noises, $\{v_k^i; k \geq 1\}$, $i = 1, 2$, with the following probability distributions

$$P[v_k^1 = -8] = \frac{1}{8}, \quad P\left[v_k^1 = \frac{8}{7}\right] = \frac{7}{8}, \quad \forall k \geq 1,$$

$$P[v_k^2 = 1] = \frac{15}{18}, \quad P[v_k^2 = -3] = \frac{2}{18}, \quad P[v_k^2 = -9] = \frac{1}{18}, \quad \forall k \geq 1.$$

To analyze the performance of the proposed estimators, the linear and quadratic filtering error variances have been calculated for different values of p^1 and p^2 . Such variances show insignificant variation from the 5th iteration onwards and, consequently, only the error variances at a specific iteration are considered; in Figure 1 the linear and quadratic filtering error variances at $k = 50$ are displayed versus p^1 (for constant values of p^2) and, in Figure 2, these variances are shown versus p^2 (for constant values of p^1). From these figures it is gathered that, as the probability that the signal is present in the observations at both sensors increases, the error variances are smaller and, hence, better estimations are obtained. Moreover, for all the values of p^1 and p^2 , the error variances corresponding to the quadratic filter are less than those of the linear filter, which means that the estimation accuracy of the quadratic filter is superior to that of the linear one.

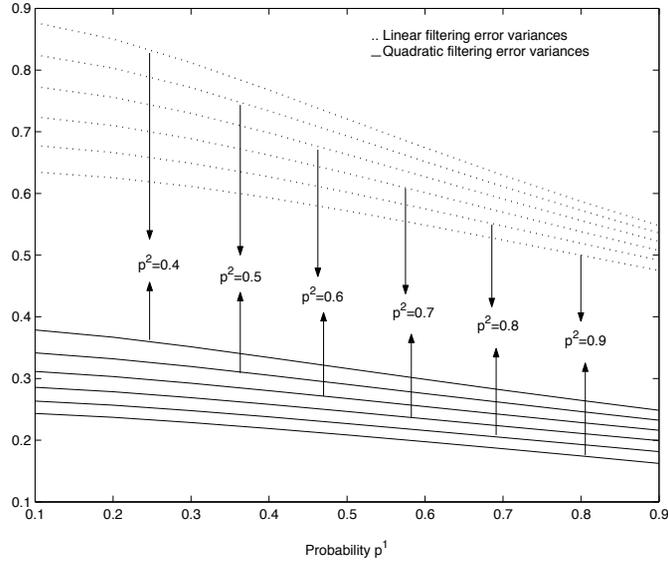


FIGURE 1. Linear and quadratic filtering error variances versus p^1 with $p^2 = 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$.

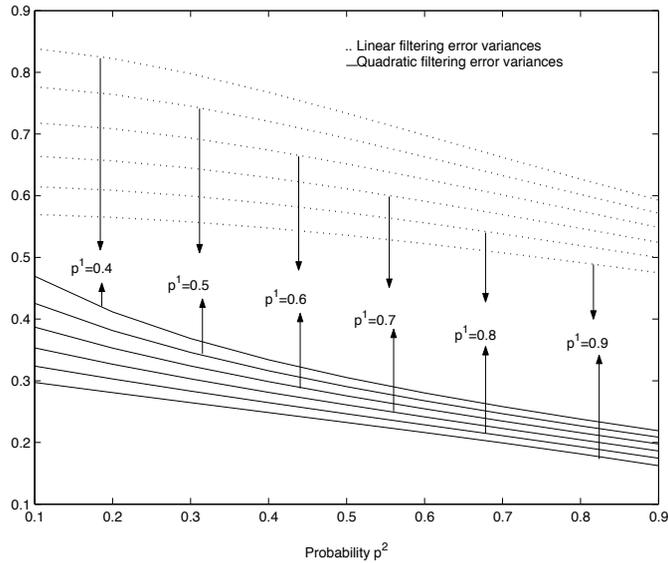


FIGURE 2. Linear and quadratic filtering error variances versus p^2 with $p^1 = 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$.

Acknowledgments: This work is partially supported by the *Ministerio de Ciencia e Innovación* and the *Junta de Andalucía* through projects MTM2008-05567 and P07-FQM-02701, respectively.

References

- De Santis, A., Germani, A., and Raimondi, M. (1995). Optimal quadratic filtering of linear discrete-time non-Gaussian systems, *IEEE Transactions on Automatic Control*, **AC-40**, 1274–1278.
- Houkpevi, F.O., and Yaz, E.E. (2007). Minimum variance linear state estimators for multiple sensors with different failure rates. *Automatica*, **43**, 1274–1280.
- Jiménez-López, J.D., Linares-Pérez, J., Nakamori, S., Caballero-Águila, R., and Hermoso-Carazo, A. (2008). Signal estimation based on covariance information from observations featuring correlated uncertainty and coming from multiple sensors. *Signal Processing*, **88**, 2998–3006.

Modelling trends in digit preference patterns

Carlo G. Camarda¹, Paul H. C. Eilers² and Jutta Gampe¹

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: A two-dimensional generalization of a penalized composite link model is presented to model latent distributions with digit preference, where the strength of the misreporting pattern can vary over time. A general preference pattern is superimposed on a series of smooth latent densities, and this pattern is modulated for each measurement occasion. Smoothness of the latent distributions is enforced by a difference penalty on neighbouring coefficients. An L_1 -ridge regression is used for the common misreporting pattern, and an additional weighted least-squares regression extracts the modulating vector. The BIC is used to optimize the smoothing parameters. We present a simulation study and an application for demonstrating the performance of our model and its practical characteristics.

Keywords: Composite Link Model; Digit preference; L_1 penalty; Penalized likelihood.

1 Introduction

Digit preference is the tendency to round measurements or other observations to pleasing digits. If several measurement occasions are available, the strength of the preference pattern may vary over time (e.g. due to gain in experience or better instruments), while its shape may be unchanged. For instance, mortality data commonly present specific misreporting patterns for ages-at-death, which mostly improve gradually over time due to more accurate vital registration, but which may also deteriorate in times of crisis. Figure 2, top-left shows ages-at-death for Spain during the period 1920–1940, taken from the Human Mortality Database (2009).

To model such structures, a two-dimensional approach has to be employed. In this paper we present a general model that allows to estimate the common misreporting pattern, its development over a second dimension (usually time) as well as the smooth latent densities devoid of misreporting.

2 Modelling digit preference in one dimension

Digit preference was modelled by Camarda et al. (2008), when only one measurement occasion is considered: actual data are assumed to be re-

alizations from a Poisson distribution with a composed mean, $\boldsymbol{\mu} = \mathbf{C}\boldsymbol{\gamma}$, and a smooth latent distribution $\boldsymbol{\gamma}$. The matrix \mathbf{C} embodies the misreporting probabilities p_{ik} and allows each count to be redistributed to the immediately neighbouring categories. The composite link model (Thompson and Baker, 1981) is thus used as a suitable framework. By defining $\check{x}_{ik} = \sum_j c_{ij} x_{jk} \gamma_j / \mu_i$, the iteratively reweighted least squares (IRWLS) algorithm can be generalized:

$$(\check{\mathbf{X}}' \tilde{\mathbf{W}} \check{\mathbf{X}} + \mathbf{P}) \tilde{\boldsymbol{\beta}} = \check{\mathbf{X}}' \tilde{\mathbf{W}} \tilde{\mathbf{z}}, \quad (1)$$

where $\tilde{\mathbf{W}} = \text{diag}(\tilde{\boldsymbol{\mu}})$, $\tilde{\mathbf{z}} = \tilde{\mathbf{W}}^{-1}(\mathbf{y} - \tilde{\boldsymbol{\mu}}) + \check{\mathbf{X}} \tilde{\boldsymbol{\beta}}$ and \mathbf{P} measures the roughness of the vector $\boldsymbol{\gamma}$ with differences of order d , weighted by a positive regularization parameter (Eilers, 2007).

The numerous misreporting probabilities in the vector \mathbf{p} are estimated by a constrained weighted least-squares regression within the IRWLS algorithm. To make the estimation feasible an L_1 -penalty is introduced (Tibshirani, 1996), which allows to select only the p_{ik} that exhibit the strongest effects. If \mathbf{p} denotes the probabilities concatenated into a vector, from the structure of \mathbf{C} we can write:

$$\boldsymbol{\mu} = \mathbf{C}\boldsymbol{\gamma} = \boldsymbol{\gamma} + \boldsymbol{\Gamma}\mathbf{p},$$

where $\boldsymbol{\Gamma}$ is the associated model matrix. Since $\mathbf{y} \sim \text{Poisson}(\boldsymbol{\mu})$, we therefore approximate $(\mathbf{y} - \boldsymbol{\gamma})$ as

$$(\mathbf{y} - \boldsymbol{\gamma}) \approx N(\boldsymbol{\Gamma}\mathbf{p}, \text{diag}(\boldsymbol{\mu})). \quad (2)$$

Consequently, the following penalized weighted least-squares system can be solved iteratively:

$$(\boldsymbol{\Gamma}' \tilde{\mathbf{V}} \boldsymbol{\Gamma} + \mathbf{Q}) \tilde{\mathbf{p}} = \boldsymbol{\Gamma}' \tilde{\mathbf{V}} (\mathbf{y} - \tilde{\boldsymbol{\gamma}}), \quad (3)$$

where $\tilde{\mathbf{V}} = \text{diag}(1/\tilde{\boldsymbol{\mu}})$ and $\mathbf{Q} = \kappa \text{diag}(1/|\mathbf{p}|)$. The size of misreporting proportions p_{ik} is tuned by the smoothing parameter κ .

3 Modelling the temporal trend

Generalizing this model to a two-dimensional setting, we assume a series of latent distributions, γ_{ij} , where $i = 1, \dots, I$ and $j = 1, \dots, J$ index measurement values and occasions, respectively. Smoothness is assumed both for the individual distribution, but also between adjacent measurement occasions. We also assume the same misreporting pattern, which, however, may be more or less pronounced at each occasion j . This is expressed by a vector $\mathbf{g} = (g_j)_j$ acting multiplicatively on the composition matrix \mathbf{C} .

3.1 A two-dimensional penalized CLM

A generalization of the IRWLS presented in equation (1) requires a different composition matrix. Including the modulating factors $(g_j)_j$, the two-dimensional composition matrix $\check{\mathbf{C}}$ therefore is:

$$\check{\mathbf{C}} = [\check{c}_{ik,j}] = \mathbf{I} + \text{diag}(\mathbf{g}) \otimes \begin{pmatrix} -p_{21} & 0 & 0 & 0 & \cdots & 0 \\ p_{21} & -p_{32} & 0 & \cdots & & \vdots \\ 0 & p_{32} & -p_{43} & 0 & \cdots & \vdots \\ 0 & 0 & p_{43} & \ddots & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & -p_{I,I-1} & 0 \\ 0 & \cdots & \cdots & 0 & p_{I,I-1} & 0 \end{pmatrix}.$$

Again, we do not consider covariates, so the model matrix \mathbf{X} will simply be the $I \times J$ identity matrix. In this way the modified model matrix $\check{\mathbf{X}}$ becomes a block diagonal matrix:

$$\check{\mathbf{X}} = \text{blockdiag}[\check{\mathbf{X}}_1, \check{\mathbf{X}}_2, \dots, \check{\mathbf{X}}_j, \dots, \check{\mathbf{X}}_J]$$

where

$$\check{\mathbf{X}}_j = \begin{pmatrix} (1 - g_j p_{21}) \cdot \frac{\gamma_{1,j}}{\mu_{1,j}} & 0 & 0 & 0 & 0 \\ g_j p_{21} \cdot \frac{\gamma_{1,j}}{\mu_{2,j}} & (1 - g_j p_{32}) \cdot \frac{\gamma_{2,j}}{\mu_{2,j}} & 0 & \cdots & \vdots \\ 0 & g_j p_{32} \cdot \frac{\gamma_{2,j}}{\mu_{3,j}} & (1 - g_j p_{43}) \cdot \frac{\gamma_{3,j}}{\mu_{3,j}} & 0 & \vdots \\ 0 & 0 & g_j p_{43} \cdot \frac{\gamma_{3,j}}{\mu_{4,j}} & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \frac{\gamma_{I,j}}{\mu_{I,j}} \end{pmatrix}$$

The system of equations (1) can be thus directly employed with a different roughness penalty:

$$\mathbf{P} = \lambda_I \mathbf{I}_I \otimes \mathbf{D}'_I \mathbf{D}_I + \lambda_J \mathbf{D}'_J \mathbf{D}_J \otimes \mathbf{I}_J, \quad (4)$$

where λ_I and λ_J are the smoothing parameters used over the two dimensions (Currie et al., 2004).

3.2 Finding the common misreporting pattern

The common misreporting probabilities are estimated using equation (3). The model matrix $\mathbf{\Gamma}$ is adapted for a two-dimensional setting and it takes

the following form:

$$\mathbf{\Gamma} = \begin{pmatrix} \mathbf{\Gamma}_1 \\ \mathbf{\Gamma}_2 \\ \dots \\ \mathbf{\Gamma}_j \\ \dots \\ \mathbf{\Gamma}_J \end{pmatrix}$$

where

$$\mathbf{\Gamma}_j = \begin{pmatrix} -g_j\gamma_{1,j} & 0 & 0 & \dots & \dots & \dots & \dots \\ g_j\gamma_{1,j} & -g_j\gamma_{2,j} & 0 & \dots & \dots & \dots & \dots \\ 0 & g_j\gamma_{2,j} & -g_j\gamma_{3,j} & \ddots & \dots & \dots & \dots \\ \vdots & 0 & g_j\gamma_{3,j} & \ddots & \dots & \dots & \dots \\ \vdots & \vdots & 0 & \ddots & -g_j\gamma_{i,j} & \dots & \dots \\ \vdots & \vdots & \vdots & \dots & g_j\gamma_{i,j} & \dots & -g_j\gamma_{I-1,j} \\ \vdots & \vdots & \vdots & \dots & \dots & \dots & g_j\gamma_{I-1,j} \end{pmatrix}.$$

Again the L_1 -penalty allow to extract the misreporting probabilities and the smoothing parameter κ control the size of p_{ik} shrinking the less important close to zero.

3.3 The temporal trend

For the scaling vector \mathbf{g} we use a weighted least-squares regression. Though possible, we do not assume smoothness for the temporal changes of the misreporting pattern. Using the approximation in (2), we solve the following system of equations for each j :

$$\boldsymbol{\theta}'_j \check{\mathbf{V}}_j \boldsymbol{\theta}_j \tilde{g}_j = \boldsymbol{\theta}'_j \check{\mathbf{V}}_j (\mathbf{y}_j - \boldsymbol{\gamma}_j), \quad (5)$$

where $\check{\mathbf{V}}_j = \text{diag}(1/\boldsymbol{\mu}_j)$ and $\boldsymbol{\theta}_{ij} = -p_{i+1,i}\gamma_{i,j} + p_{i,i-1}\gamma_{i-1,j}$.

The parameterization in equation (5) is not unique, because it is invariant with respect to any linear combination of \mathbf{g} and \mathbf{p} . It has been sufficient to constrain the maximum of \mathbf{g} to be equal to 1.

3.4 Optimal smoothing

The estimating equations (1), (3) and (5) depend on the combination of the three smoothing parameters λ_I , λ_J and κ . To optimize these parameters we minimize Bayesian Information Criterion (BIC), where the effective dimension is the sum of the three model components. In formula:

$$\text{BIC}(\lambda_I, \lambda_J, \kappa) = \text{Dev}(\mathbf{y}|\boldsymbol{\mu}) + \ln(IJ) [\text{ED}_1 + \text{ED}_2 + \text{ED}_3]. \quad (6)$$

$\text{Dev}(\mathbf{y}|\boldsymbol{\mu})$ is the deviance of the Poisson model. We chose the effective dimension as the sum of the three model components, i.e. ED_1 denotes the effective dimension of the two-dimensional penalized CLM, ED_2 refers to the L_1 -ridge regression for the common misreporting pattern and ED_3 is equal to the length of modulating vector. Specifically, we have

$$\text{ED}_1 = \text{trace}\{\check{\mathbf{X}}(\check{\mathbf{X}}'\hat{\mathbf{W}}\check{\mathbf{X}} + \mathbf{P})^{-1}(\check{\mathbf{X}}'\hat{\mathbf{W}})\},$$

$$\text{ED}_2 = \text{trace}\{\boldsymbol{\Gamma}(\boldsymbol{\Gamma}'\hat{\mathbf{V}}\boldsymbol{\Gamma} + \mathbf{Q})^{-1}\boldsymbol{\Gamma}'\hat{\mathbf{V}}\} \quad \text{and} \quad \text{ED}_3 = J.$$

Instead of a plain grid-search over a complete range of values, we efficiently explore a three-dimensional space of $[\lambda_I, \lambda_J, \kappa]$, optimizing each smoothing parameter in turn, moving at most one grid step up or down.

4 Simulation and Application

To demonstrate the performance of the model, we applied it to simulated scenario (see Figure 1). In this scenario, digits 5 and 20 attracted additional observations from neighbouring categories. Moreover, we assumed a specific trend for the misreporting pattern over j .

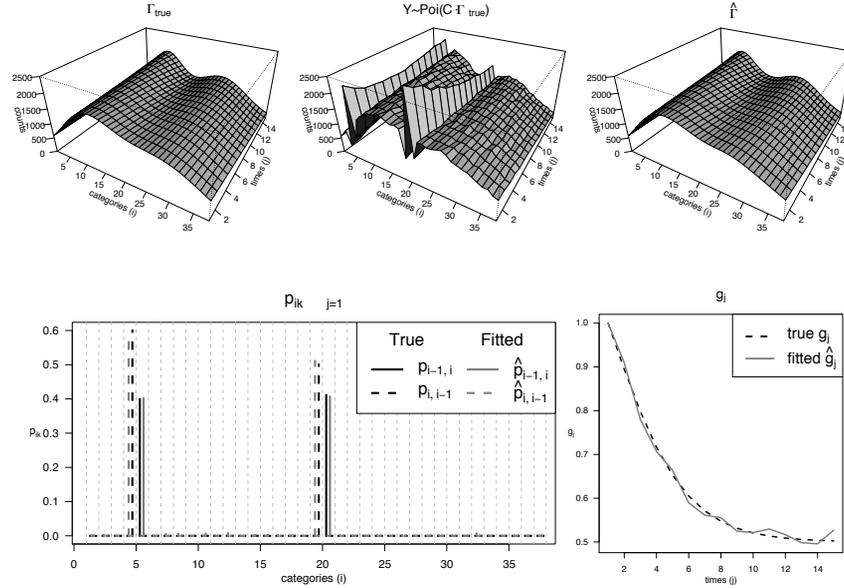


FIGURE 1. Simulated data. True values (top-left), raw data (top-central) and estimates (top-right). True misreporting probabilities and estimates (bottom-left). Scaling vector modulating the misreporting pattern (bottom-right).

The top panels in Figure 1 show the true latent surface (left), a possible simulated \mathbf{Y} (central) and the estimated surface from such a simulation (right). The bottom-left panel summarizes true and estimated misreporting probabilities when $j = 1$. Such a pattern is then modulated by g_j as shown in the bottom-right panel.

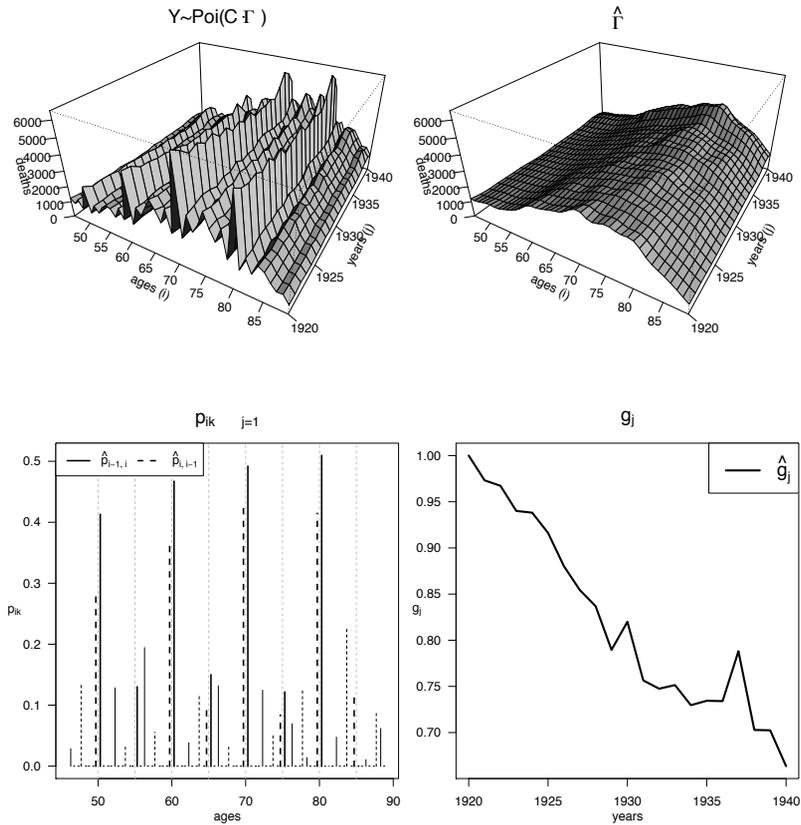


FIGURE 2. Spanish females, ages-at-death: observed counts (top-left) and estimated true latent distribution (top-right). Fitted misreporting probabilities (bottom-left). Scaling vector modulating the preference pattern (bottom-right).

If we apply the model to the Spanish mortality data, we obtain the results shown in Figure 2. The top panel shows the actually recorded (left) and smoothly fitted death counts (right). The seemingly under-smoothing behaviour of the fitted surface is due to differences in cohort sizes. Such differences are explicitly not include in our model and they thus may produce diagonal effects on the surface.

The bottom-left image in Figure 2 summarizes the estimated misreporting probabilities when $j = 1$. Factors g_j (bottom-right) modulate this pattern over years. Ages that end in 5 or 10 clearly attract observations, the latter ones showing the strongest effects. Moreover, people tend to underestimate their age, i.e. ages that are multiples of 10 show a slightly higher tendency to receive counts from their respective right neighbours. Finally, there is a relative peak of \mathbf{g} in 1937, which can be interpreted as a clear effect of the Spanish Civil War on the data collection system.

5 Discussion

In this paper we present a method for coping with digit preference in a two-dimensional setting. Data are assumed to be indirect observations from a series of latent densities combined with a misreporting pattern. Such pattern is common for every density, though modulated for each measurement occasion.

Smoothness is the only assumption made about the series of latent densities and a generalization of the penalized composite link model is considered. To ensure smoothness and reduce the effective dimension, a roughness penalty is used on neighbouring categories over both dimensions.

Transferring observations from any adjacent digits is allowed in the common misreporting pattern. An L_1 -penalty guarantees the feasibility of the estimation and it selects only the misreporting probabilities that exhibit the strongest effects. In such flexible setting, different level of misreporting is possible between end-digits.

The simulation study and the application on actual mortality data have shown that our model can properly capture the latent densities devoid of digit preference. Moreover, the fitted misreporting pattern may be intrinsically interesting and the resulting temporal trend allows additional interpretations of the digit preference developments.

In many demographic data, digit preference are manifest in both death counts and population exposure. One can envision the possibility to simultaneously extract misreporting pattern from both series of densities. This approach would allow further analysis free of any age heaping and it would also avoid cohort effects due to different cohort sizes.

More general patterns of misreporting can be easily incorporate in the presented framework and an augmented version of the model matrix $\mathbf{\Gamma}$ would allow exchanges between digits that are more than one category apart. Though such extension will enormously increase the number of misreporting probabilities, early results have shown that L_1 -ridge regression is still adequate for selecting p_{ik} .

Challenging areas for further research would include the computational aspects of the model. Two-dimensional penalized composite link models and generalized linear array models (Currie et al., 2006) share several features and the understanding of such similarities would enhance the IRWLS

algorithm for the smooth latent densities. Boosting algorithms for regularization can also be adopted to improve the L_1 -ridge regression component.

References

- Camarda, C. G., Eilers, P. H. C. and Gampe J. (2008). Modelling General Patterns of Digit Preference. *Statistical Modelling*, **8**, 385-401.
- Currie, I. D., Durban, M. and Eilers, P. H. C. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, **4**, 279-298.
- Currie, I. D., Durban, M. and Eilers, P. H. C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of Royal Statistical Society. Series B*, **68**, 259-280.
- Eilers, P. H. C. (2007). Ill-posed Problems with Counts, the Composite Link Model, and Penalized Likelihood. *Statistical Modelling*, **7**, 239-254.
- Human Mortality Database (2009). 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.
- Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, **58**, 267-288.

CHASSIS - Nonparametric Bayesian Estimates of Gravitational Potential and Phase Space Density Function

Dalia Chakrabarty¹

¹ School of Physics & Astronomy, University of Nottingham, Nottingham NG7 2RD, U.K.

Abstract: We discuss the Bayesian non-parametric algorithm CHASSIS that estimates phase space density function and gravitational potential of a relaxed astrophysical system, using incomplete, 1-component velocity information of system members. The unknown functions are constrained via a maximum likelihood approach and the code implements an MCMC optimizer. We also discuss the development of a Bayesian significance test for implementation in non-parametric contexts and implement it to analyze kinematic data of distinct types of galactic members in an example galaxy.

Keywords: Bayes theorem, Bayesian significance test, Astrophysical applications.

1 Introduction

A dynamical system comprises phase space W , the points (\mathbf{w}) in which are the possible states of the system, along with a rule that determines the state at any time t , given the initial state $\mathbf{w}(t_0)$ (Alligood et. al 1996, Katok et. al 1997). Here $\mathbf{w} = \mathbf{x} + \mathbf{v}$, \mathbf{x} is the 3-D spatial vector and $\mathbf{v} = \dot{\mathbf{x}}$. For a system with n particles, the mapping of the t coordinate into the $3n$ -D “configuration space”, or $\mathbf{x} : \mathbb{R} \rightarrow \mathbb{R}^{3n}$, represents the motion of the system. The distribution of phase space points is given by the *phase space density function* $f(\mathbf{x}, \mathbf{v}, t)$. The evolution of a sample of phase space points, drawn from $f(\cdot)$ at time t_0 , is deterministic given the function $\mathbf{g} : \mathbb{R}^{3n} \times \mathbb{R}^{3n} \times \mathbb{R} \rightarrow \mathbb{R}^{3n}$ such that $\ddot{\mathbf{x}} = \mathbf{g}(\mathbf{x}, \mathbf{v}, t)$. This expresses Newton’s equation of motion (Arnold 1978, Jordan et. al 1999). Here \mathbf{g} is defined if the system potential $\Phi(\mathbf{x})$ is known. Thus, the complete characterization of a dynamical system would require knowledge of $f(\cdot)$ and $\Phi(\cdot)$.

We attempt determination of $f(\cdot)$ and $\Phi(\cdot)$ for non-relativistic gravitational systems such as galaxies. The ulterior aim is to constrain the distribution of the total matter density $\rho(\mathbf{x})$ in such systems where Poisson equation states $\Phi(\mathbf{x}) = -4\pi\mathbf{G}\nabla^2\rho(\mathbf{x})$, (G is the Universal Gravitational constant). This is a non-trivial exercise given that the major contributor to the total mass

density is dark matter while the readily available astronomical observations are photometric or luminous in nature. Since astrophysical theory does not provide any analytical relationship between photometric information and dark matter content, such data is considered irrelevant for our purpose. Instead, we rely upon measurements of the line-of-sight (LOS) component of velocities of individual galactic members: v_3 , say (in addition to the positions of the individual galactic members with respect to the system centre, on the plane-of-the-sky). Such 1-D kinematic data is typically sparse and incomplete, with $\lesssim 100$ data points available for a nearby galaxy.

Conventionally, in astrophysical literature, the determination of the total mass distribution - even when undertaken independently of photometric information - typically suffers from deprojection uncertainties, dependence on details of binning of the sparse data, assumptions of smooth and analytical $f(\cdot)$ and $\Phi(\cdot)$ and ultimately on the implementation of goodness-of-fit estimators which are spuriously inflated in the presence of (large) inhomogeneous errors of measurement - very much the reality of astronomical kinematic measurements (Bissantz & Munk 2001). The assumption of smooth $f(\cdot)$ and $\Phi(\cdot)$ is misrepresentative since galaxies are complex objects; the high degree of phase space non-linearity manifests even in a local patch around the Solar position in our galaxy (Chakrabarty 2007, Chakrabarty & Sideris 2008). Thus, a non-parametric approach is more welcome.

This is offered by the Bayesian non-parametric algorithm CHASSIS that determines $f(\cdot)$ and $\Phi(\cdot)$ via a maximum-likelihood approach and uses measured 1-D velocity data as input (Chakrabarty & Saha 2001). The relevant optimization is MCMC in nature (Metropolis-Hastings). Actually, CHASSIS determines $\rho(\cdot)$ and the potential $\Phi(\cdot)$ is obtained from it, using Poisson equation. This helps avoid running into problems with negative densities which can arise if the potential is updated during an iterative step.

2 Methodology

Dynamical theory tells us that $f(\cdot)$ can be a function of \mathbf{w} through the integrals of motion only, i.e.

$$f = f(K_i[\mathbf{w}]) \quad \text{where} \quad \dot{K}_i = 0 \quad \text{for} \quad i = 1, 2, \dots \quad (1)$$

Given the limited size of the available data, we realize that we need to constrict the degrees of freedom in our work. In fact, we consider $i = 1, 2 : K_1 = E$ and $K_2 = L_3$, where the particle energy is $E = \Phi + v^2/2$, $v^2 = \sum_{i=1}^3 v_i^2$ and the LOS component of the angular momentum vector is $L_3 = x_1 v_2 - x_2 v_1$. Also, we assume radial symmetry, i.e. $\Phi = \Phi(r)$ and $\rho = \rho(r)$ where $r^2 = \sum x_i^2$.

We estimate $\Pr(f(\cdot), \rho(\cdot)|data)$ and estimate this using Bayes theorem. The only priors that we can impose upon $f(\cdot)$ or $\rho(\cdot)$, are:

$$f(E, L_3) \geq 0 \quad \text{and} \quad \left. \frac{\partial f(E, L_3)}{\partial E} \right|_{L_3} < 0, \quad \forall E, L_3$$

$$\rho(r) \geq 0 \quad \text{and} \quad \frac{d\rho}{dr} < 0, \quad \forall r. \quad (2)$$

Other than such monotonicity & positivity conditions, the sought functions are completely free-form. Assuming $f = f(E, L_3) \implies$ phase space is anisotropic. However, if we consider $f(\cdot)$ to depend only on E , it will imply an isotropic dependence of $f(\cdot)$ on \mathbf{x} and \mathbf{v} .

To connect a trial $f(\cdot)$ - at a trial $\Phi(\cdot)$ - to the k^{th} line in the N_{data} sized data sample (x_1^k, x_2^k, v_3^k) , we project $f(\cdot)$ into the space of observables. This projection of $f(\cdot)$ involves Φ via E :

$$\nu_k(x_1^k, x_2^k, v_3^k) = \int f[E(\Phi(\mathbf{x}), \mathbf{v}), L_3(x_1, x_2, v_1, v_2)] dx_3 dv_1 dv_2 \quad (3)$$

The likelihood function is defined as $\mathcal{L} = \sum_{k=1}^{N_{data}} \log \nu_k$.

2.1 Histograms

In order to compute the projection integral in Equation 3 for a given data point, we discretize the $E - L_3$ space and the relevant radial range. Then we confer histogram-like structures to the unknown $f(\cdot)$ and $\Phi(\cdot)$, over the $E - L_3$ and r bins respectively. Thus,

$$f(E_i, L_3^j) = \alpha_{ij} \quad \text{for} \quad E \in [E_{i-1}, E_i], \quad L_3 \in [L_3^{j-1}, L_3^j], \\ i = 1, \dots, N_{eng}, \quad j = 1, \dots, N_{L_3}. \quad (4)$$

Here α_{ij} is a constant for a given i, j , Similarly,

$$\rho(r_i) = \rho_i \quad \text{for} \quad r \in [r_{i-1}, r_i], \quad i = 1, \dots, N_r. \quad (5)$$

Under the assumption of stratification of mass in spherical shells, $\Phi(r)$ is recovered from the current r -histogram. The projection integral in Equation 3, is carried out over the $i - j^{th}$ $E - L_3$ cell, $\forall i, j$. This requires the mapping of the integral $I_{ij} = \int_{i,j} dv_1 dv_2 dx_3$ into the $E - L_3$ space. The integral over the $v_1 - v_2$ space is analytical $\forall i, j$ while that over x_3 is performed numerically using the lower and upper limits of 0 and x_3^{ij-UP} where for the k^{th} line in the data set, $x_3^{ij-UP} = \sqrt{(r_E^{ij})^2 - (x_1^k)^2 - (x_2^k)^2}$, with r_E^{ij} being the solution to the equation $E_i = v_3^2/2 + (L_3^j)^2/2(r_E^{ij})^2 + \Phi(r_E^{ij})$. The $E - L_3$ -histogram and r -histogram are tweaked in shape and size in every iterative step. For example, the ansatz adopted for the updating the shape of ρ in the $q + 1^{th}$ step is:

$$\rho_i^{q+1} = \rho_{i+1}^q + (\rho_i^q - \rho_{i+1}^q) \exp\left(\frac{\mathcal{R}}{s_1}\right) \quad (6)$$

The r -histogram is then scaled by a factor of $\exp(\mathcal{R}/s_2)$. Here \mathcal{R} is a random number in $[-0.5, 0.5]$ and s_1 and s_2 are experimentally optimized

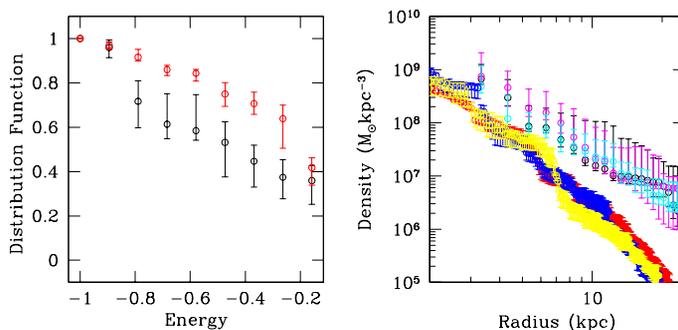


FIGURE 1. *Right:* $\rho(r)$ recovered from runs of CHASSIS done with 2 velocity data sets ($SPNe$ & SGC) of 2 different types of members (PNe & GC) that live in the same galaxy. The results of the $SPNe$ -runs are shown in red, yellow and blue while $\rho(r)$ obtained by using SGC are in black, magenta and cyan. For a given data, the runs vary from each other in the choice of the initial seed. As is apparent from the figure, starting with different seeds results in density profiles that are consistent with each other, as long as the same kinematic tracer sample is used in the analysis. *Left:* isotropic (normalized) $f(E)$ recovered from $SPNe$ -run (in red) and from SGC -run (in black). The dependence of recovered galactic gravitational mass density distribution, on the member type, is demonstrated.

scales. Similarly, the $E-L_3$ -histogram is updated in shape and subsequently normalized.

CHASSIS uses the Metropolis-Hastings optimization (Metropolis et. al 1953, Hastings 1970, Chib & Greenberg 1995). It is envisaged (and borne in test runs) that the likelihood structure is multimodal and non-linear. Given this state of affairs, we implement highly dispersed seeds to initiate several chains (Gelman & Rubin 1992) and have found it useful to use simulated annealing on a single chain. The details of the implemented cooling schedule are determined experimentally. The search for a better alternative to the currently used optimization is work underway, though exploration of implementations of RJMCMC (Richardson & Green, 1997) is noted to be too expensive to suggest inclusion.

2.2 How CHASSIS works

The estimates of $f(\cdot)$ and $\Phi(\cdot)$ follow from the posterior distribution: $\Pr(\alpha_{11}, \dots, \alpha_{N_{eng} N_{L_3}}, \rho_1, \dots, \rho_{N_r} | data)$, as modeled within the MCMC approach. At the beginning of every iterative step, the $E-L_3$ -histogram and r -histogram are updated and the ensuing $f(\cdot)$ -structure is projected into the space of observables, at the updated potential structure, at each line of data. The resulting projection defines the likelihood, the global maxima of which is sought.

2.3 Assuming Isotropy

In practice, the paucity of available data is so severe that it becomes difficult to constrain the total galactic matter density distribution within useful uncertainty levels ($\pm 1\text{-}\sigma$ errors of $\lesssim 50\%$). This scenario prompts us to reduce a d.o.f. in the calculations, so that we assume: $f = f(E)$, i.e. we assume the phase space to be *isotropic in \mathbf{v} and \mathbf{x}* . This implies that the number of L_3 bins reduces to 1.

3 Test of Isotropy

The simplifying assumption of isotropy naturally throws up the question of the validity of the same, given the data. Addressing this question is basically an exercise in hypothesis testing. In Chakrabarty & Saha (2001), such was addressed via p -value estimates. However, sensitivity of the p -value measure on sample size obfuscates interpretation - a Bayesian formalism is a better alternative, eg. Fully Bayesian Significance Test or FBST (Stern & Pereira 1999, Pereira et. al 2008). The null hypothesis that we aim to test is that isotropy is a good assumption to make in CHASSIS. In other words, we test if the phase space density from which the input data are drawn, is isotropic, i.e.:

$$H_0 : \hat{f} = \Psi[E(\sum_i v_i^2/2 + \Phi(r))] \quad (7)$$

Here \hat{f} is the phase space density function from which the input kinematic data are drawn. Ψ is some function: $\Psi > 0$ for $E < 0$ and $\Psi = 0$ otherwise. Thus H_0 is sharp.

Within FBST, let the null hypothesis be $H_0 : \theta = \theta_0$, ($\theta \in \Theta$, say, with θ assumed distributed uniformly in Θ -space). Then the evidence in favor of H_0 is $1 - ev$ where ev is the evidence value such that:

$$ev = 1 - \Pr(\theta \in T | data), \quad \text{where } T = \{\theta : \Pr(\theta | data) > \Pr(\theta^* | H_0)\}. \quad (8)$$

Here the probability of recovering θ , given the data is $\Pr(\theta | data)$ and θ^* is the point in Θ space, satisfying H_0 , that maximizes this posterior probability. T is the tangential set. Thus, FBST involves identification of θ^* , followed by integration over T .

A non-parametric implementation of FBST is developed. In our work $\theta \equiv \{\alpha_{i1}, \rho_k\}$, $i = 1, \dots, N_{eng}$, $k = 1, \dots, N_r$. Here, Θ is the space of all $E - L_3$ -histograms and r -histograms.

Our MCMC optimizer, upon convergence, identifies a $\rho(r)$ and an isotropic $f(E)$, with a $1\text{-}\sigma$ error band on each. The data used to achieve these solutions do not necessarily satisfy our null. From this achieved $f(E)$, we draw a random sample of $\{x_1, x_2, v_3\}$, i.e. the observable quantities. These generated data abide by H_0 since they were drawn from the isotropic phase space density $f(E)$. These data are used to initiate N more runs (say) of

CHASSIS. A likelihood is achieved at the end of each step in each of these runs. Of these runs, let during the n^{th} run, at the end of the m^{th} step, the highest likelihood \mathcal{L}^* is achieved. Then

$$\{\alpha_{i1}^{mn}, \rho_k^{mn}\} \equiv \theta^*, \quad i = 1, \dots, N_{eng}, k = 1, \dots, N_r. \quad (9)$$

We compare \mathcal{L}^* with the likelihood values obtained at all steps within all other undertaken runs, and find (say) that in A steps out of a total of B cases, the achieved likelihood exceeds the identified \mathcal{L}^* . Then,

$$\Pr(\theta \in T | data) = \frac{A}{B} = 1 - ev. \quad (10)$$

Currently, we are looking into the objective qualification of the “small”ness of a recovered ev in terms of minimization of the loss function (Pereira et. al 2008, Madurga et. al 2001).

4 Application

The current version of our non-parametric implementation of FBST has been implemented in a comparative sense to judge the goodness of the assumption of isotropy in two different data sets (S_{PNe} & S_{GC}) obtained by measuring velocities of two distinct types of galactic members (PNe & GC). While S_{PNe} indicates $1 - ev \approx 0.6$, $S_{GC} \implies 1 - ev \approx 0.95$. Expectedly, the $f(E)$ recovered from runs done with the two data are different. We also find that $\rho(r)$ obtained from the two distinct data are distinct (see Fig.1). This demonstrates the risk involved in the extraction of the galactic mass distribution from kinematic data of one type of members only.

References

- Alligood, K. T., Sauer, T., & Yorke, J. A. (1996) *Chaos: an introduction to dynamical systems*, Springer.
- Arnold, V. I. (1978) *Mathematical Methods of Classical Mechanics*, Springer.
- Bissantz, N., & Munk, A. (2001) *Astronomy & Astrophysics*, **376**, 735.
- Chakrabarty, D., & Sideris, I. V. (2008), *Astronomy & Astrophysics*, **488**, 161.
- Chakrabarty, D. (2007), *Astronomy & Astrophysics*, **467**, 145.
- Chakrabarty, D. & Portegies Zwart, S. (2004), *Astronomical J.*, **128**, 1046.
- Chakrabarty, D. & Saha, P. (2001), *Astronomical J.*, **122**, 232.
- Chib, S. & Greenberg, E. (1995), *American Statistician*, **49**, 327.
- Gelman, A. & Rubin, D. B. (1992), *Statistical Science*, **7**, 457.

- Hastings, W. K. 1970, *Biometrika*, **57**, 97.
- Jordan, D., & Smith, P. (1999) *Nonlinear Ordinary Differential Equations: An Introduction to Dynamical Systems*, Oxford Univ Press.
- Katok, A., Hasselblatt, B., & Mendoza, L. (1997) *Introduction to the modern theory of dynamical systems*, Cambridge Univ. Press.
- Madruga, R. M., Esteves, I. G. & Wechsler, S. (2001) *Test*, **10**, 291.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A., & Teller, H. (1953), *Jl. of Chemical Physics*, **21**, 1087.
- Pereira, C. A. de B., Stern, J. M. & Wechsler, S., (2008), *Bayesian Analysis*, **3**, 79.
- Pereira, C. A. de B. & Stern, J. M. (1999), *Entropy*, **1**, 99.
- Richardson, P. J. & Green, S. (1997) *Jl. of the Royal Statistical Society. Series B*, **59**, 731.

Fast Bayesian Functional Data Analysis: Application to basal body temperature data

James M. Ciera¹, Bruno Scarpa¹ and David B. Dunson²

¹ Department of Statistical Science, University of Padova. Via Cesare Battisti 241, 35121 Padova, Italy. Email: ciera@stat.unipd.it, scarpa@stat.unipd.it

² Department of Statistical Science, Duke University. Box 90251, Durham, NC 27708-0251 USA. Email: dunson@stat.duke.edu

Abstract: Functional data collected repeatedly from many subjects is commonly characterized with unequal cycle lengths and unequally-spaced measurements. Borrowing information across subjects is crucial while analyzing such data. However, while accounting for these limitations, most Markov Chain Monte Carlo (MCMC) based methods are slow a factor that raises a practical motivation for fast methods. Motivated by methods proposed in the machine learning literature, we present an application of a fast approximate Bayes functional data analysis method to rapidly estimate individual-specific functions based on basal body temperature data.

Keywords: Basis functions; MAP estimates; Relevance vector machine; Sparsity.

1 Introduction

In many clinical studies, data is collected repeatedly from many subjects over a period of time. Using massive datasets, physicians require fast automated tools to estimate data trajectories and predict clinically important events for a current patient. For example, in reproductive studies, trajectories of hormonal level or daily basal body temperature (bbt) among women can help to identify or predict early pregnancy loss and occurrence of the ovulation day (Bigelow and Dunson, 2008). Borrowing of information among the subjects is crucial when observations are sparse and the interest is in prediction. Therefore, there is a need for fast algorithms for estimating functional trajectories while borrowing information from other patients about the shape and location of features in the function.

1.1 Motivating problem

Our research is motivated by the bbt data from European fecundability study (Colombo and Masarotto, 2000). The study enrolled women aged between 18 and 40 years, were not taking hormonal medications or drugs

affecting fertility, and had no known impairment of fecundity. The participants kept daily records of basal body temperature, and recorded the days during which intercourse and menstrual bleeding occurred. Ovulation days were estimated for each menstrual cycle under study using the last day of hypothermia prior to the post-ovulatory rise in basal body temperature. For more details about the study protocol, refer to Colombo and Masarotto (2000). We consider the daily bbt measurements from women that contributed temperature measurements from at least one menstrual cycle. The data is characterized with unequal cycle lengths and unequally-spaced measurements causing problems in estimating the bbt curves. Thus, estimation of accurate and smooth curves is based on borrowing information flexibly across cycles.

1.2 Multi-Task Relevant Vector Machine

Functional data analysis (FDA) can be used to estimate trajectories but relies on large number of basis functions (Ramsay and Silverman, 1997). Bayesian methods can be implemented to estimate basis coefficients but the posterior sampling is based on slow Markov Chain Monte Carlo (MCMC) algorithms. This raises a practical motivation for fast approximate Bayes approaches that bypass MCMC while maintaining some of the benefits of a Bayesian analysis. In this article, we approximate the bbt projectiles using Multi-Task Relevant Vector machine (MT-RVM) method (Ji, *et al*, 2008) an extension of Relevant Vector machine (RVM) method propose by Tipping, (2001). RVM is a fast Bayesian method based on Empirical Bayes methodology and penalizes the basis coefficients through a scale mixture of normals prior, which is carefully-chosen so that maximum a posteriori (MAP) estimates of many of the coefficients are zero. This provides a natural mechanism in selection of the basis functions leading to a sparse models that is fast to compute.

2 Methods

We consider observations from the i^{th} woman with response vector $\mathbf{y}_i = (y_{i1}, \dots, y_{iT_i})'$ consisting of bbt measurements and covariate vector $\mathbf{z}_i = (z_{i1}, \dots, z_{iT_i})'$ representing observation day. A functional model is represented as,

$$y_{it} = f_i(z_{it}) + \epsilon_{it}, \quad \epsilon_{it} \sim N(0, \sigma_\epsilon^2), \quad t = 1, \dots, T_i, \quad i = 1, \dots, N. \quad (1)$$

where $f_i(\cdot)$ is a smooth function for subject i and ϵ_{it} is a measurement error. The smoothing function can be described as a linear combination of basis functions $f_i(z_{it}) = \sum_{j=1}^M \beta_{ij} \varphi_j(z_{it}) = \mathbf{x}_{it}' \boldsymbol{\beta}_i$ where $\mathbf{x}_{it} = (x_{it1}, \dots, x_{itM})'$ are the values of the basis functions at z_{it} , parameter β_{ij} is the coefficient for the j^{th} basis function $\varphi_j(\cdot)$ and $\boldsymbol{\beta}_i = (\beta_{i1}, \dots, \beta_{iM})'$.

The priors are $\beta_{ij} \sim N(0, \alpha_j^{-1})$, $\sigma_\epsilon^{-2} \sim \text{Gamma}(a, b)$ and $\alpha_j \sim \text{Gamma}(c, d)$ where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_M)'$ and σ_ϵ^{-2} are computed from the data and shared among the subjects. Computation of the posterior density is based on the conditional distribution $p(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_\epsilon^{-2} | \mathbf{Y}) = p(\boldsymbol{\alpha}, \sigma_\epsilon^{-2} | \mathbf{Y}) \prod_{i=1}^N p(\boldsymbol{\beta}_i | \mathbf{y}_i, \boldsymbol{\alpha}, \sigma_\epsilon^{-2})$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_N)$. The posterior for $\boldsymbol{\beta}_i$ is,

$$p(\boldsymbol{\beta}_i | \mathbf{Y}, \boldsymbol{\alpha}, \sigma_\epsilon^{-2}) = N(\boldsymbol{\beta}_i; \hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i), \quad (2)$$

where $\hat{\boldsymbol{\mu}}_i = \sigma_\epsilon^{-2} \hat{\boldsymbol{\Sigma}}_i \mathbf{X}_i' \mathbf{y}_i$ and $\hat{\boldsymbol{\Sigma}}_i = (\mathbf{A} + \sigma_\epsilon^{-2} \mathbf{X}_i' \mathbf{X}_i)^{-1}$ such that $\mathbf{A} = \text{diag}\{\alpha_1, \dots, \alpha_M\}$ and $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{iM})'$.

Expressing $p(\boldsymbol{\alpha}, \sigma_\epsilon^{-2} | \mathbf{Y})$ is difficult analytically and the MAP estimates for $\boldsymbol{\alpha}$ and σ_ϵ^{-2} are computed from the marginal likelihood $p(\mathbf{Y} | \boldsymbol{\alpha}, \sigma_\epsilon^{-2})$, obtained after integrating out $\boldsymbol{\beta}_i$ from $p(\mathbf{Y} | \boldsymbol{\beta}_i, \sigma_\epsilon^{-2})$. This results to a normal density function $N(\mathbf{y}_i; \mathbf{0}, \mathbf{C}_i)$ where the covariance matrix $\mathbf{C}_i = \sigma_\epsilon^2 \mathbf{I}_{T_i} + \sum_{j=1}^M \alpha_j^{-1} \mathbf{x}_{ij} \mathbf{x}_{ij}'$. The expressions for the estimates of $\boldsymbol{\alpha}$ and σ_ϵ^{-2} are obtained after differentiating the expression for $p(\mathbf{Y} | \boldsymbol{\alpha}, \sigma_\epsilon^{-2})$ with respect to parameters $\boldsymbol{\alpha}$ and σ_ϵ^{-2} respectively and equating the resulting expressions to zero. This results to

$$\hat{\alpha}_j = \frac{N}{\sum_{i=1}^N \mu_{ij}^2 + \boldsymbol{\Sigma}_{i,jj}}, \quad \text{and} \quad \hat{\sigma}_\epsilon^{-2} = \frac{\sum_{i=1}^N \|\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\mu}_i\|^2}{\sum_{i=1}^N (T_i - M - \boldsymbol{\alpha}^{-1} \boldsymbol{\Sigma}_{i,jj})}. \quad (3)$$

Estimates for $\boldsymbol{\alpha}$ and σ_ϵ^{-2} are inserted into $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\Sigma}}_i$ in equation (2) leading to an interactive procedure alternating between estimation of the parameters in equation (2) and (3) respectively. However, when the dimensions of \mathbf{X}_i is large, inverting the $M \times M$ matrix in equation (2) becomes problematic. The computation process becomes slow and inefficient prompting the need for a fast and efficient method to compute $\boldsymbol{\alpha}$.

A fast approach to compute the elements of $\boldsymbol{\alpha}$ is done sequentially. This is based on the dependence of $\ell(\boldsymbol{\alpha}, \sigma_\epsilon^{-2}) = \log p(\mathbf{Y} | \boldsymbol{\alpha}, \sigma_\epsilon^{-2})$ upon the k^{th} element of $\boldsymbol{\alpha}$ leading to the decomposing of $\ell(\boldsymbol{\alpha}, \sigma_\epsilon^{-2})$ into two parts -one with and without the k^{th} element of $\boldsymbol{\alpha}$. The solutions from the resulting score equations are infeasible to express analytically except for $\alpha_k = \infty$. To avoid complexities, we assume that $\alpha_k \ll s_{ik}$, leading to an approximate estimate

$$\hat{\alpha}_k \cong \begin{cases} \frac{N}{\sum_{i=1}^N (q_{ik}^2 - s_{ik}) / s_{ik}^2} & \text{if } \sum_{i=1}^N \frac{(q_{ik}^2 - s_{ik})}{s_{ik}^2} > 0, \\ \infty & \text{otherwise.} \end{cases} \quad (4)$$

where $s_{ik} = \mathbf{x}_{ik}' \mathbf{C}_{i,-k}^{-1} \mathbf{x}_{ik}$, $q_{ik} = \mathbf{x}_{ik}' \mathbf{C}_{i,-k}^{-1} \mathbf{y}_i$ and $\mathbf{C}_{i,-k}$ is the component of \mathbf{C}_i^{-1} without the contribution of the k^{th} basis function. The estimate for $\hat{\sigma}_\epsilon^{-2}$ is as expressed in equation (3). For a justification of this type of approximation, refer to Ji, et al. (2008).

We first start with an empty model and select the basis function that has the largest impact on the log-likelihood $\ell(\boldsymbol{\alpha}, \sigma_\epsilon^{-2})$. The subsequent steps on

selection of the remaining basis functions involves three operations on \mathbf{X}_{ik} . These are; addition, deletion or updating $\hat{\mu}_{ik}$ operations. Addition occurs when $\sum_{i=1}^N \frac{(q_{ik}^2 - s_{ik})}{s_{ik}^2} > 0$ and \mathbf{X}_{ik} is not in the model, while an update occurs when \mathbf{X}_{ik} is already in the model and $\sum_{i=1}^N \frac{(q_{ik}^2 - s_{ik})}{s_{ik}^2} > 0$. We delete the basis function \mathbf{X}_{ik} from the model when $\sum_{i=1}^N \frac{(q_{ik}^2 - s_{ik})}{s_{ik}^2} < 0$.

3 Results

We present results from the bbt data and also show the performance of the MT-RVM method relative to a classical MCMC based method as the number of observations increases. The O’Sullivan-type Penalized splines method (Wand and Ormerod, 2008) is used to generate the basis functions.

3.1 Application to the bbt data

We considered data from 500 women and select cycles that had identified ovulation day. Each woman contributed data from their first menstrual cycle in the study. We generated the basis function and estimated the curves using the MT-RVM and the classical Bayesian method. Figure 1 presents estimated curves for one cycle using the two methods where the continuous and dotted curves represent the curves generated by the MT-RVM and MCMC based methods respectively. The gray region is the credible band for the classical Bayesian method.

The MCMC based curve is generated by estimating 22 non-zero basis coefficients while the approximation by the MT-RVM method is based on estimates for 4 non-zero basis coefficients. On time factor, the MCMC based method takes an average of 19.35 seconds to generate a curve while the MT-RVM method takes an average of 0.09 seconds to select the relevant basis coefficients and generate a curve.

3.2 Effects of adding more observations

To assess the performance of the MT-RVM with the increase in the number of observations per cycle, we generated 30 biphasic curves that mimic the shape of the bbt curve. The curves were generated using a sine function,

$$y_{it} = v_i + \rho_i z_{it} \sin(10z_{it} - r_i) + \epsilon_{it}, \quad t = 1, 2, \dots, 27, \quad i = 1, 2, \dots, 30,$$

where covariate $z_{it} \sim \text{unif}(0, 1)$, while $v_i \sim \text{unif}(-1, 1)$ and $r_i \sim \text{unif}(-1, 1)$ are the vertical and horizontal shift parameters and $\rho_i \sim \text{unif}(0.5, 1.5)$ controls the amplitude of the curves. Each curve had 27 observations and we generated the basis functions using the method used in the previous section. The computation of the basis coefficients was based on the two

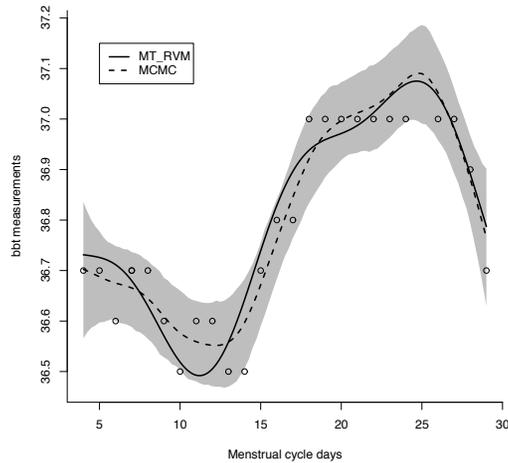


FIGURE 1. Estimated bbt curves using the MT-RVM and MCMC methods.

methods -an MCMC based and the MT-RVM methods. To compare the estimation of the basis coefficients from the two methods as the number of observations increases, we computed Reconstructive Error

$$RE_l = \frac{1}{N} \sum_{i=1}^N \frac{\|\beta_i^{RVM} - \beta_i^{MCMC}\|}{\|\beta_i^{RVM}\|}.$$

After the computation of the coefficients for the initial model, we subsequently simulated additional 200 observations for each cycle. After each increment, we re-computed the basis functions \mathbf{X}_i for the new data, computed the basis coefficients using the two methods and then computed the reconstructive error value. We plotted the RE_l against the number of observations (l) as shown in figure 2. It is evident that the increase of the number of observations leads to a gradual decreases in the reconstructive error but the decrease trend reaches to a constant value after 150 observations. However, the reconstructive error curve remains constant at a non-zero RE value since most of the non-relevant basis coefficients from the MCMC based method are non-zero while their corresponding basis coefficients from the MT-RVM method are zero.

3.3 Concluding note

The paper demonstrates the use of a fast Empirical Bayes method as an alternative to computer intensive methods that rely on MCMC. The method

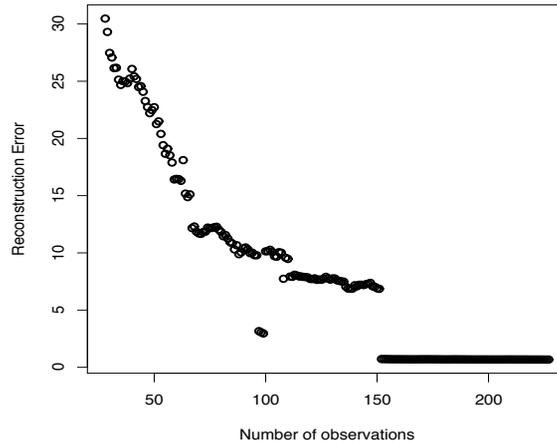


FIGURE 2. A plot for RE against the number of observations.

is fast and can be used to rapidly approximate curves for massive datasets. In our subsequent work, we have extended the MT-RVM method to cover linear mixed effect (LME) models and hierarchical data where a woman can have data from multiple cycles.

References

- Bigelow, J.L., and Dunson, D.B. (2008). Bayesian adaptive regression splines for hierarchical data. *Biometrics*, **63**,724-732.
- Colombo, B. and Masarotto, G. (2000). Daily fecundability: First results from a new data base. *Demographic Research*, **3**, 5.
- Ji, S., Dunson, D.B. and Carin, L. (2008). Multi-task compressive sensing. *IEEE Transactions on Signal Processing*, to appear.
- Ramsay, J.O. and Silverman, B.W. (1997). *Functional data analysis*. New York: Springer Verlag.
- Tipping, M.E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, **1**, 211-244.
- Wand, M.P. and Ormerod, J.T. (2008). On O’Sullivan penalised splines and semiparametric regression. *Australian and New Zealand Journal of Statistics*, **50**, 179-198.

Multi Edge Graphs for Multivariate Markov Chains

Roberto Colombi¹ and Sabrina Giordano²

¹ Università di Bergamo - Italy, colombi@unibg.it;

² Università della Calabria - Italy, sabrina.giordano@unical.it

Abstract: The aim of this paper is to provide a graphical representation of the dynamic relations among the marginal processes of a first order multivariate Markov chain. We show how to read Granger-noncausal and contemporaneous independence relations off a particular type of graph, the multi edge graph, when directed and bi-directed edges are missing. Multivariate logistic models for transition probabilities are associated with multi edge graphs which encode the interdependencies of interest. A categorical multivariate time series concerning access to web servers is used to illustrate the relevance of the proposed models.

Keywords: Graphical models; Granger noncausality; Marginal interactions.

1 Introduction

The identification of dynamic relations among variables, simultaneously observed over time, is an interesting task in many areas. Two types of dependence relations in multivariate time series models are basically relevant: the dependence of the present of a subset of variables on the past of all the variables, and the contemporaneous dependencies among variables at any time point that cannot be ruled out by conditioning on the past.

In this paper we address the use of graphical models for the analysis of the dynamic relations among the marginal processes of a time-homogeneous first order multivariate Markov chain. We employ a multi edge graph whose nodes represent the univariate marginal processes of the Markov chain and whose directed and bi-directed edges describe the dependence structure among them. The approach, adopted here, enables us to interpret the lack of directed edges as Granger-noncausal relationships (see Chamberlain, 1982) while missing bi-directed edges are used to visualize the contemporaneous independence relations between the marginal processes of the chain.

The transition probabilities of the multivariate Markov chain are required to obey this set of Markov properties implied by such a graph and we prove how this problem is equivalent to testing linear constraints on appropriate interaction parameters. A similar graphical approach was used by Eichler (2007) to describe the dynamic structure of multivariate autoregressive processes.

2 Basic Notation

Given a set of integers $\mathcal{V} = \{1, \dots, q\}$, let $\mathbf{A}_{\mathcal{V}} = \{A_{\mathcal{V}}(t) : t \in \mathcal{Z}\} = \{A_j(t) : t \in \mathcal{Z}, j \in \mathcal{V}\}$ be a time-homogeneous first order multivariate Markov chain (MMC) in a discrete time interval $\mathcal{Z} = \{0, 1, 2, \dots\}$. For all $t \in \mathcal{Z}$, $A_{\mathcal{V}}(t) = \{A_j(t) : j \in \mathcal{V}\}$ is a discrete random vector with each element $A_j(t)$ taking on values in a finite set $\mathcal{A}_j = \{a_{j1}, \dots, a_{js_j}\}$, $j \in \mathcal{V}$. For every $\mathcal{S} \subset \mathcal{V}$, a marginal process of the chain is represented by $\mathbf{A}_{\mathcal{S}} = \{A_{\mathcal{S}}(t) : t \in \mathcal{Z}\}$ where $A_{\mathcal{S}}(t) = \{A_j(t) : j \in \mathcal{S}\}$. When $\mathcal{S} = \{j\}$, the univariate marginal process is indicated as \mathbf{A}_j , $j \in \mathcal{V}$. Moreover, we use the notation of conditional independence $X \perp\!\!\!\perp Y | W$ to signify that the random variables X and Y are independent once the value of a third variable W is given.

3 Graphical Multivariate Markov Chains

Our aim is to provide a graphical representation of the dynamic relations among the component processes of a MMC. The basic problem here lies in finding a graph that encodes G-noncausal and contemporaneous independence statements. This is achieved by a particular graphical structure that we call *Multi Edge Graph* (ME graph). In the ME graph $G = (\mathcal{V}, \mathcal{E})$, with the node set \mathcal{V} and the edge set \mathcal{E} , there must exist a one-to-one correspondence between the nodes $j \in \mathcal{V}$ and the univariate marginal processes \mathbf{A}_j , $j \in \mathcal{V}$, of the MMC $\mathbf{A}_{\mathcal{V}}$. A pair of nodes $i, k \in \mathcal{V}$ of the ME graph may be joined by the directed edges $i \rightarrow k$, $i \leftarrow k$, and by the bi-directed edge $i \leftrightarrow k$, denoted also by (i, k) , $[i, k)$ and $[i, k]$, respectively. Each pair of distinct nodes $i, k \in \mathcal{V}$ can be connected by up to all the three types of edges. For each single node $i \in \mathcal{V}$, the bi-directed edge $[i, i]$ is implicitly introduced and the directed edge (i, i) may or may not be present. If $(i, k) \in \mathcal{E}$ then i is a *parent* of k and k is a *child* of i . The set of parents of the node i is denoted by $Pa(i) = \{j \in \mathcal{V} : (j, i) \in \mathcal{E}\}$. Moreover, when $[i, k) \in \mathcal{E}$ the nodes i, k are *neighbors*. Thus, the set of neighbors of i is $Nb(i) = \{j \in \mathcal{V} : [i, j) \in \mathcal{E}\}$. Note that the generic node i is neighbor of itself ($i \in Nb(i)$) and may also be parent and child of itself. More generally, $Pa(\mathcal{S})$ and $Nb(\mathcal{S})$ are the collection of parents and neighbors of nodes in \mathcal{S} , for every non-empty subset \mathcal{S} of \mathcal{V} (see the wide ranging literature on graphical models for basic concepts).

ME graphs obey Markov properties which associate sets of G-noncausality and contemporaneous independence restrictions with missing directed and bi-directed edges, respectively. In particular, missing bi-directed arrows lead to independencies concerning marginal processes at the same point in time; missing directed edges, instead, refer to independencies which involve marginal processes at two consecutive instants.

We can now introduce the key definition of our work.

Definition 1. (Graphical MMC). *A multivariate Markov chain is graphical with respect to an ME graph $G = (\mathcal{V}, \mathcal{E})$ if and only if its transition probabilities satisfy the following conditional independencies for all $t \in \mathcal{Z} \setminus \{0\}$*

$$A_{\mathcal{S}}(t) \perp\!\!\!\perp A_{\mathcal{V} \setminus Pa(\mathcal{S})}(t-1) | A_{Pa(\mathcal{S})}(t-1) \quad \forall \mathcal{S} \in \mathcal{P}(\mathcal{V}) \quad (1)$$

$$A_{\mathcal{S}}(t) \perp\!\!\!\perp A_{\mathcal{V} \setminus Nb(\mathcal{S})}(t) | A_{\mathcal{V}}(t-1) \quad \forall \mathcal{S} \in \mathcal{P}(\mathcal{V}). \quad (2)$$

Condition (1) is equivalent, for all $t \in \mathcal{Z} \setminus \{0\}$, to the two statements

$$A_{\mathcal{S}}(t) \perp\!\!\!\perp A_{\mathcal{V} \setminus Pa(\mathcal{S}) \cup \mathcal{S}}(t) | A_{Pa(\mathcal{S}) \cup \mathcal{S}}(t-1) \quad \forall \mathcal{S} \in \mathcal{P}(\mathcal{V}) \quad (3)$$

$$A_{\mathcal{S}}(t) \perp\!\!\!\perp A_{\mathcal{S} \setminus Pa(\mathcal{S})}(t-1) | A_{Pa(\mathcal{S})}(t-1) \quad \forall \mathcal{S} \in \mathcal{P}(\mathcal{V}). \quad (4)$$

In the context of first order MMC, condition (3) corresponds to the classical notion of Granger noncausality (Colombi and Giordano, 2009). In particular, for all $\mathcal{S}' \in \mathcal{P}(\mathcal{V})$ where $\mathcal{S}' \cap (Pa(\mathcal{S}) \cup \mathcal{S}) = \emptyset$, (3) implies $A_{\mathcal{S}}(t) \perp\!\!\!\perp A_{\mathcal{S}'}(t-1) | A_{\mathcal{V} \setminus \mathcal{S}'}(t-1)$ which means that the most recent past of $\mathbf{A}_{\mathcal{S}'}$ is irrelevant to predict $\mathbf{A}_{\mathcal{S}}$ once the most recent past of $\mathbf{A}_{\mathcal{V} \setminus \mathcal{S}'}$ is known. In this case, it is usual to say that $\mathbf{A}_{\mathcal{S}'}$ does not G-cause $\mathbf{A}_{\mathcal{S}}$ with respect to $\mathbf{A}_{\mathcal{V}}$. Thus, $Pa(\mathcal{S})$ identifies the maximal marginal process $\mathbf{A}_{\mathcal{V} \setminus Pa(\mathcal{S}) \cup \mathcal{S}}$ which does not G-cause $\mathbf{A}_{\mathcal{S}}$ with respect to $\mathbf{A}_{\mathcal{V}}$.

Statement (4) concerns the independence of a process from its own past, that is, whenever $\mathcal{S} \setminus Pa(\mathcal{S}) \neq \emptyset$, it describes the case of variables at the current time-point which are not affected by their immediate past.

Henceforth, we will refer to (1) with the term *Granger noncausality condition* saying that $\mathbf{A}_{\mathcal{S}}$ is not G-caused by $\mathbf{A}_{\mathcal{V} \setminus Pa(\mathcal{S})}$ with respect to $\mathbf{A}_{\mathcal{V}}$, and we will use the shorthand notation $\mathbf{A}_{\mathcal{V} \setminus Pa(\mathcal{S})} \not\rightarrow \mathbf{A}_{\mathcal{S}}$.

On the other hand, condition (2) is a restriction on marginal transition probabilities because it does not involve the marginal processes $\mathbf{A}_j : j \in Nb(\mathcal{S}) \setminus \mathcal{S}$, at time t , and more precisely it states that the transition probabilities must satisfy the bi-directed Markov property (Richardson, 2003) with respect to the graph obtained by removing the directed edges from G . Here, we will refer to (2) with the term *contemporaneous independence condition* using a shorthand notation $\mathbf{A}_{\mathcal{S}} \leftrightarrow \mathbf{A}_{\mathcal{V} \setminus Nb(\mathcal{S})}$, and we will say that $\mathbf{A}_{\mathcal{S}}$ and $\mathbf{A}_{\mathcal{V} \setminus Nb(\mathcal{S})}$ are contemporaneously independent (with respect to $\mathbf{A}_{\mathcal{V}}$ is assumed implicitly).

The above Definition suggests that the lack of a directed edge from node i to k , ($i, k \in \mathcal{V}$), is equivalent to the independence of the present of the univariate marginal process \mathbf{A}_k from the immediate past of \mathbf{A}_i given the most recent past of the marginal process $\mathbf{A}_{\mathcal{V} \setminus \{i\}}$, that is, for all $t \in \mathcal{Z} \setminus \{0\}$

$$(i, k] \notin \mathcal{E} \iff A_k(t) \perp\!\!\!\perp A_i(t-1) | A_{\mathcal{V} \setminus \{i\}}(t-1). \quad (5)$$

From Definition 1, moreover, we deduce that a missing bi-directed arrow between i and k is equivalent to stating that the corresponding marginal

FIGURE 1. *Example of a multi edge graph*

processes are contemporaneously independent given the recent past of the MMC, that is, for all $t \in \mathcal{Z} \setminus \{0\}$

$$[i, k] \notin \mathcal{E} \iff A_i(t) \perp\!\!\!\perp A_k(t) | A_{\mathcal{V}}(t-1). \quad (6)$$

The conditional independencies (5) and (6) are interpretable in terms of pairwise Granger noncausality and contemporaneous independence conditions, respectively. The more general noncausal and contemporaneous independence statements of Definition 1 are needed because the pairwise restrictions, associated with missing edges, are in general not sufficiently strong for encoding all the Granger noncausal relations and contemporaneous independence properties among the components of an MMC. This follows from the fact that the composition property (Eichler, 2007) does not hold in the context of multivariate Markov chains. For example, the condition

$$A_{\mathcal{S}}(t) \perp\!\!\!\perp A_i(t-1) | A_{\mathcal{V} \setminus \{i\}}(t-1),$$

is not equivalent to

$$A_k(t) \perp\!\!\!\perp A_i(t-1) | A_{\mathcal{V} \setminus \{i\}}(t-1), \forall k \in \mathcal{S},$$

which means that the G-noncausality for the joint process $\mathbf{A}_{\mathcal{S}}$ is not equivalent to the G-noncausality for all the univariate processes \mathbf{A}_k , $k \in \mathcal{S}$.

Example. The graph in Figure 1 displays the contemporaneous independence relation $\mathbf{A}_{12} \leftrightarrow \mathbf{A}_3$ and the G-noncausal restrictions: $\mathbf{A}_1 \nrightarrow \mathbf{A}_3$; $\mathbf{A}_{13} \nrightarrow \mathbf{A}_2$; $\mathbf{A}_3 \nrightarrow \mathbf{A}_1$; $\mathbf{A}_1 \nrightarrow \mathbf{A}_{23}$; $\mathbf{A}_3 \nrightarrow \mathbf{A}_{12}$. The pairwise G-noncausality conditions associated to the missing directed edges in graph 1 are: $\mathbf{A}_1 \nrightarrow \mathbf{A}_2$; $\mathbf{A}_1 \nrightarrow \mathbf{A}_3$; $\mathbf{A}_3 \nrightarrow \mathbf{A}_1$; $\mathbf{A}_3 \nrightarrow \mathbf{A}_2$. Note that, for example, relations $\mathbf{A}_3 \nrightarrow \mathbf{A}_1$; $\mathbf{A}_3 \nrightarrow \mathbf{A}_2$ are not equivalent to $\mathbf{A}_3 \nrightarrow \mathbf{A}_{12}$ and this shows how the pairwise conditions do not imply the more general causal restriction (1). Similarly, the pairwise contemporaneous independence conditions $\mathbf{A}_1 \leftrightarrow \mathbf{A}_3$, $\mathbf{A}_2 \leftrightarrow \mathbf{A}_3$ do not imply $\mathbf{A}_{12} \leftrightarrow \mathbf{A}_3$ given by (2).

4 A Multivariate Logistic Model for Transition Probabilities

We remind the reader that $\mathcal{I} = \times_{j \in \mathcal{V}} \mathcal{A}_j$ is the joint state space. The time homogeneous joint transition probabilities are denoted by $p(\mathbf{i} | \mathbf{i}')$, for every

pair of states $\mathbf{i} \in \mathcal{I}$, $\mathbf{i}' \in \mathcal{I}$. Given a vector $\mathbf{i} = (i_1, i_2, \dots, i_q)' \in \mathcal{I}$, if $\mathcal{M} \subset \mathcal{V}$ then $\mathbf{i}_{\mathcal{M}}$ denotes the vector with components $i_j : j \in \mathcal{M}$. Given a state $\mathbf{i}' \in \mathcal{I}$, for the transition probabilities $p(\mathbf{i}|\mathbf{i}')$, $\mathbf{i} \in \mathcal{I}$, we adopt the Glonek-McCullagh (1995) multivariate logistic model whose marginal interaction parameters are denoted by $\eta^P(\mathbf{i}_P|\mathbf{i}')$, for every non empty subset P of \mathcal{V} and for every $\mathbf{i}_P \in \times_{j \in P} \mathcal{A}_j$. The Glonek-McCullagh baseline interactions $\eta^P(\mathbf{i}_P|\mathbf{i}')$ are given by the following contrasts of logarithms of marginal transition probabilities $p(\mathbf{i}_P|\mathbf{i}')$ from the state \mathbf{i}' to one of the states in $\times_{j \in P} \mathcal{A}_j$

$$\eta^P(\mathbf{i}_P|\mathbf{i}') = \sum_{\mathcal{K} \subseteq P} (-1)^{|P \setminus \mathcal{K}|} \log p((\mathbf{i}_{\mathcal{K}}, \mathbf{i}_{P \setminus \mathcal{K}}^*|\mathbf{i}')). \quad (7)$$

We can observe that the Glonek-McCullagh interactions are not log-linear parameters because they are not contrasts of logarithms of the joint transition probabilities $p(\mathbf{i}|\mathbf{i}')$.

To model the dependence of the transition probabilities on the conditioning states $\mathbf{i}' \in \mathcal{I}$, we adopt the usual factorial expansion of Glonek-McCullagh marginal interactions $\eta^P(\mathbf{i}_P|\mathbf{i}') = \sum_{Q \subseteq \mathcal{V}} \theta^{P,Q}(\mathbf{i}_P|\mathbf{i}'_Q)$.

The Möbius inversion theorem (Lauritzen, 1996) ensures that

$$\theta^{P,Q}(\mathbf{i}_P|\mathbf{i}'_Q) = \sum_{\mathcal{H} \subseteq Q} (-1)^{|Q \setminus \mathcal{H}|} \eta^P(\mathbf{i}_P|(\mathbf{i}'_{\mathcal{H}}, \mathbf{i}'_{\mathcal{V} \setminus \mathcal{H}}^*)). \quad (8)$$

Thus, it easily turns out that the transition probabilities are parameterized by the interactions $\theta^{P,Q}(\mathbf{i}_P|\mathbf{i}'_Q)$, $P \subseteq \mathcal{V}, P \neq \emptyset, Q \subseteq \mathcal{V}, \mathbf{i}_P \in \times_{j \in P} \mathcal{A}_j, \mathbf{i}'_Q \in \times_{j \in Q} \mathcal{A}_j$. The proof of next proposition follows from classical results on the logistic regression and a result by Lupparelli et al. (2009).

Proposition 1. *For an MMC with strictly positive time homogeneous transition probabilities, it holds that: (i) the Granger noncausality condition (1) is true if and only if all interactions $\theta^{P,Q}(\mathbf{i}_P|\mathbf{i}'_Q)$ are equal to zero, for $P \subseteq \mathcal{V}, P \neq \emptyset, Q \not\subseteq Pa(P)$; and (ii) the contemporaneous independence condition (2) is equivalent to the zero constraints on the interactions $\theta^{P,Q}(\mathbf{i}_P|\mathbf{i}'_Q)$ for all P that are not connected sets in the bi-directed graph obtained by removing every directed edge from the ME graph G .*

The Proposition states that the requirements (1), (2) for a graphical MMC correspond to simple linear constraints on the $\theta^{P,Q}(\mathbf{i}_P|\mathbf{i}'_Q)$ parameters and thus testing the hypotheses (1), (2) becomes a standard parametric problem.

The expression (7) of the Glonek-McCullagh interactions in terms of baseline log-linear contrasts of marginal transition probabilities is not necessarily the most convenient. In fact, when the interpretation of the non null interaction parameters is of interest and the $\mathcal{A}_j, j \in \mathcal{V}$, are ordered sets, more general types of interactions can be used as shown by Bartolucci et al. (2007). From Bartolucci et al., it can be proved that the set

TABLE 1. Hypothesis Testing

Hyp	Missing edges	G^2	df	p -value	Relations
1	[1, 2], [1, 3], [2, 3]	80.37	32	0.00	$\mathbf{A}_1 \leftrightarrow \mathbf{A}_2 \leftrightarrow \mathbf{A}_3$
2	[1, 2], [1, 3]	22.32	24	0.56	$\mathbf{A}_1 \leftrightarrow \mathbf{A}_{2,3}$
3	(1, 2), (1, 3), (2, 1), (3, 1)	23.81	18	0.16	$\mathbf{A}_1 \leftrightarrow \mathbf{A}_{2,3}, \mathbf{A}_{2,3} \rightarrow \mathbf{A}_1$
4	(2, 1), (2, 3), (1, 2), (3, 2)	32.47	18	0.02	$\mathbf{A}_2 \rightarrow \mathbf{A}_{1,3}, \mathbf{A}_{1,3} \rightarrow \mathbf{A}_2$
5	(3, 1), (3, 2), (1, 3), (2, 3)	24.95	18	0.13	$\mathbf{A}_3 \rightarrow \mathbf{A}_{1,2}, \mathbf{A}_{1,2} \rightarrow \mathbf{A}_3$
6	[1, 2], [1, 3], (1, 2), (1, 3), (2, 1), (3, 1)	45.09	42	0.34	$\mathbf{A}_1 \leftrightarrow \mathbf{A}_{2,3}, \mathbf{A}_1 \rightarrow \mathbf{A}_{2,3}, \mathbf{A}_{2,3} \rightarrow \mathbf{A}_1$
7	[1, 2], [1, 3], (3, 1), (3, 2), (1, 3), (2, 3)	45.30	38	0.19	$\mathbf{A}_1 \leftrightarrow \mathbf{A}_{2,3}, \mathbf{A}_3 \rightarrow \mathbf{A}_{1,2}, \mathbf{A}_{1,2} \rightarrow \mathbf{A}_3$
8	[1, 2], [1, 3], [2, 3], (1, 2), (1, 3), (2, 1), (3, 1), (2, 3), (3, 2)	110.2	50	0.00	$\mathbf{A}_i \leftrightarrow \mathbf{A}_{j,k}, \mathbf{A}_i \rightarrow \mathbf{A}_{j,k}, \mathbf{A}_{j,k} \rightarrow \mathbf{A}_i$ $i \neq j \neq k, i, j, k = \{1, 2, 3\}$

of zero restrictions imposed on parameters $\theta^{P,Q}(i_P|i'_Q)$ can be written in the form $\mathbf{C} \ln(\mathbf{M}\boldsymbol{\pi}) = \mathbf{0}$, where $\boldsymbol{\pi}$ is the vector of all the transition probabilities and \mathbf{C} and \mathbf{M} are matrices of known constants. The procedures for the maximum likelihood estimation and hypothesis testing, developed by Lang (2005) and Cazzaro and Colombi (2009) under the assumption of Poisson-multinomial sampling and constraints $\mathbf{C} \ln(\mathbf{M}\boldsymbol{\pi}) = \mathbf{0}$, can be easily adapted to the MMC context. These procedures are implemented in the R function *Mphfit* (Lang, 2008) and in the R-package *hmmm* (Cazzaro and Colombi, 2008).

5 Example

The proposed methodology is used to analyze binary data collected every day for 6 months by an Italian mobile telephone company. The data consist of a 3-dimensional time series of the daily utilization rate level (low and high) of 3 web servers located in Rome (Italy). The data are available from the authors. The joint dynamic behavior of the series is described by a first order 3-variate Markov chain.

As the servers are all simultaneously operational, it is interesting to verify whether the status of a server depends on the utilization levels of the others on the same day, given the past use of all servers. Moreover, it is also important to ascertain whether the current working of a server is influenced by how much the others worked the day before. One answer to these questions can be attained by testing the hypotheses of G-noncausality and contemporaneous independence. This problem can be easily reduced to establishing which Markov properties of type (1) and (2) are satisfied by the transition probabilities of the web server Markov chain and to identifying the ME graph which encodes them.

To this end, various hypotheses associated with missing edges in the ME graph have been tested. Some of the results are illustrated in Table 1. Hypothesis 6 in Table 1 of no causal and contemporaneous dependence relations between the servers 1 and 2,3 is accepted. This means that yesterday's

TABLE 2. Interactions $\theta^{P,Q}$ which are null for the conditions of G-noncausality (*) and contemporaneous independence (+) encoded by the ME graph in Figure 2.

P	Q	Null	P	Q	Null	P	Q	null
1			12		+	123		+
	1			1	+		1	+
	2	*		2	+		2	+
	3	*		3	+		3	+
	12	*		12	+		12	+
	13	*		13	+		13	+
	23	*		23	+		23	+
	123	*		123	+		123	+
2			13		+			
	1	*		1	+			
	2			2	+			
	3			3	+			
	12	*		12	+			
	13	*		13	+			
	23			23	+			
	123	*		123	+			
3			23					
	1	*		1	*			
	2			2				
	3			3				
	12	*		12	*			
	13	*		13	*			
	23			23				
	123	*		123	*			

utilization levels of the servers 2 and 3 do not add helpful information to forecast the first one's operation level today and vice versa, and moreover there is no influence between the contemporaneous working of servers 1 and 2,3. The conditions under hypothesis 6 are encoded by the ME graph $G = (\mathcal{V}, \mathcal{E})$ displayed in Figure 2, with one node for each server. Therefore, the Markov chain of the web data is graphical with respect to this ME graph. Note that a directed edge from each node to itself belongs to \mathcal{E} , even if it is not drawn in Figure 2.

The missing edges of graph in Figure 2 are equivalent to a set of zero constraints on Gloneck-McCullagh interactions $\theta^{P,Q}(i_P|i_Q)$ as explained in Proposition 1. In this example, all the marginal processes have only two states, so for every P and Q there is only one Gloneck-McCullagh multivariate logistic interaction which will be denoted by $\theta^{P,Q}$. In each row of Table 2 an interaction $\theta^{P,Q}$ is identified by the sets $P \subseteq \{1, 2, 3\}$ and $Q \subseteq \{1, 2, 3\}$. The short form 123 is used to denote the set $\{1, 2, 3\}$, 12 is used to denote $\{1, 2\}$ and so on. The symbols * and + in Table 2, indicate the interactions $\theta^{P,Q}$ which have to be set to zero in order to meet the G-noncausality and the contemporaneous independence conditions which



FIGURE 2. *ME graph for web data. It encodes the G-noncausality and contemporaneous independence relations: $\mathbf{A}_1 \not\rightarrow \mathbf{A}_{2,3}$, $\mathbf{A}_{2,3} \not\rightarrow \mathbf{A}_1$; $\mathbf{A}_1 \leftrightarrow \mathbf{A}_{2,3}$.*

we can read off the ME graph in Figure 2.

References

- Bartolucci, F., Colombi, R., and Forcina, A. (2007). An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. *Statistica Sinica*, **17**, 691-711.
- Chamberlain, G. (1982). The General equivalence of Granger and Sims causality. *Econometrica*, **50**, 569-581.
- Cazzaro, M., and Colombi, R. (2008). *R package hmmm*. www.unibg.it/pers/?colombi.
- Cazzaro, M., and Colombi, R. (2009). Multinomial-Poisson models subjected to inequality constraints. *Statistical Modelling*, forthcoming.
- Colombi, R., and Giordano, S. (2009). Graphical models for multivariate Markov chains. *submitted*.
- Eichler, M. (2007). Granger causality and path diagrams for multivariate time series. *Journal of Econometrics*, **137**, 334-353.
- Gloneck, G.J.N., and McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society, Series B*, **57**, 533-546.
- Lang, J.B. (2005). Homogeneous linear predictor models for contingency tables. *Journal of the American Statistical Association*, **100**, 121-134.
- Lang, J.B. (2008). *mph.fit. An R program for maximum likelihood fitting of multinomial-Poisson homogeneous (MPH) models for contingency tables*. www.stat.uiowa.edu/~jblang/#software.
- Lauritzen, S.L. (1996). *Graphical Models*. Oxford: Clarendon Press.
- Lupparelli, M., Marchetti, G.M. and Bergsma, W.P. (2009). Parameterizations and fitting of bi-directed graph models to categorical data. *Scandinavian Journal of Statistics*, forthcoming, arXiv:0801.1440v1.
- Richardson, T. (2003). Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, **30**, 145-157.

Matched case-control data: a Bayesian partition modelling approach to mapping residual spatial variation in disease risk

Deborah A Costain

¹ Department of Mathematics and Statistics, Lancaster University, U.K. LA1 4YF

Abstract: A Bayesian partition model formulation, based upon a reversible jump MCMC scheme, is developed to provide a flexible means of modelling and mapping geo-referenced health data based upon a matched case-control design. The methodology permits spatial discontinuity and relaxes the assumptions of distance defined covariance structure. The methodology is applied to a study in to the spatial distribution of perinatal mortality. When adjusted for known spatially varying confounder Carstairs index no evidence of spatial variation in risk was evident. Lower social status was found to be indicative of an increased odds of death in this vulnerable period of life.

Keywords: Bayesian partitioning; Matched case-control data; Reversible jump MCMC; Disease mapping

1 Introduction

Methods for investigating spatial variation in disease risk continue to motivate much research; dynamics of atypical sub-regions can be used to filter resources, generate clues as to aetiology and guide further research. In general, variation due to known risk factors is not of intrinsic interest and thus the ability to accommodate known confounder's is paramount. In addition, flexibility with regard to the underlying form of the risk surface is enthused.,

Assuming individual level geo-referenced health data, confounding can be handled at the analysis stage, or alternatively, at the design stage by using the confounder's as stratifying factors for matching the controls to the cases. A consequence of the matching is that the selection process needs to be accommodated in the analysis, which in turn leads to consistency issues since the number of parameters increases with the sample size. In general, analysis of matched data proceeds on the basis of the matched conditional likelihood (Breslow and Day, 1980) in which the 'nuisance' parameters are eliminated.,

This paper presents a Bayesian partition model formulation for the analysis of matched case-control data. The work builds upon methodology devel-

oped to handle the completely randomised case-control design (Costain, 2008) a paper which outlines other key references. Features of the approach include relaxation of the customary assumption of a stationary, isotropic, covariance structure, and moreover, the ability to detect spatial discontinuity. The methodology is also generalised to handle additional 'non-matched-for' covariate information.,

Section 2 outlines the proposed model formulation and the MCMC methodology developed is outlined briefly in Section 3. In Section 4 the methodology is demonstrated on matched perinatal mortality data derived in the North-West Thames region of the UK.

2 A Bayesian Partition Model Formulation

Partition models are developed under the assumption that the region of interest, A say, can be sub-divided into a number of disjoint regions, R_j , $j = 1, \dots, k$, in which the responses, y_i , are exchangeable and derive from the same probability distribution f . In this work Voronoi tessellations (Green and Sibson, 1978) are utilised as means of partitioning due to their computational tractability and inherent flexibility. Quite simply, Voronoi tessellations are defined by a number of generating points, themselves defining the regional centres, c_j . The corresponding polygonal region, often termed *tile*, is then given by

$$R_j = \{x \in A : \|x - c_j\| < \|x - c_i\| \forall i \neq j\},$$

which is simply those points in A which are *closer* to tile centre c_j than any other tile centre. The number of tiles k and the location centres c_j are taken to be unknown.,

Now suppose we have an individually-matched case-control study, consisting of $i = 1, \dots, n$ matched groups and $j = 1, \dots, m$ elements within each group, and that group i yields the exposure vectors E_{i1}, \dots, E_{im} but it is not known which of them relates to the case and which to the controls. Following Breslow and Day, 1980, the probability that the first value E_{i1} relates to the case, $y_{i1} = 1$, and the remainder to the controls, $y_{ij} = 0$, $j = 2, \dots, m$, is given by:

$$\frac{P(y_{i1} = 1 | E_{i1}) \prod_{j=2}^m P(y_{ij} = 0 | E_{ij})}{\sum_{j=1}^m P(y_{ij} = 1 | E_{ij}) \prod_{r \neq j}^m P(y_{rj} = 0 | E_{rj})}$$

which can be re-expressed in the form:

$$\left\{ 1 + \sum_{j=2}^m \frac{P(y_{ij} = 1 | E_{ij})P(y_{i1} = 0 | E_{i1})}{P(y_{i1} = 1 | E_{i1})P(y_{ij} = 0 | E_{ij})} \right\}^{-1}. \quad (1)$$

Based upon a linear logsitic model for disease incidence:

$$\text{logit}(P(y_{ij} = 1 | E_{ij})) = \alpha_i + z'_{ij}\beta + \mu_{l(x_{ij})},$$

where the α_i 's represent the group specific intercept terms, β represents the additional unmatched-for covariate effects related to individual level covariates z_{ij} and $\mu_{l(x_{ij})}$ denotes the residual spatial height associated with region $l = 1, \dots, k$ to which the spatial location x_{ij} of individual ij belongs, and substitution in 1 the conditional probability for the n observed independent entities is given by

$$\prod_{i=1}^n \left\{ 1 + \sum_{j=2}^m \exp[(z_{ij} - z_{i1})' \beta + (\mu_{l(x_{ij})} - \mu_{l(x_{i1})})] \right\}^{-1} \quad (2)$$

which is independent of the nuisance parameters, $\alpha_1, \dots, \alpha_n$, and depends only upon the parameters of interest. Note, that the conditional likelihood is a product of multinomial probabilities.

Whilst the conditional 'matched' likelihood provides a means of consistent estimation of the exposure odds ratios of interest, the matched likelihood is a function of the *exposure differences*, $E_{ij} - E_{i1}, j = 2, \dots, m$, for each matched group i thus increasing the computational intensity. As such, analyses are based upon a Poisson likelihood formulation: conditioning on the appropriate sums.

To see this suppose that the responses are independent Poisson realisations, such that

$$y_{ij} \sim \text{Poisson}(\lambda_{ij}), \quad (3)$$

where $i = 1, \dots, n$ denotes the tuple and $j = 1, \dots, m$ the elements within tuples, and we have $N = mn$ observations in total. Without loss of generality the labelling scheme $y_{i1} = 1$ for the case in tuple i and $y_{ij} = 0, j = 2, \dots, m$, for controls is adopted and in this Poisson setting utilise a log-linear formulation:

$$\lambda_{ij} = \exp \left[\alpha_i + z'_{ij}\beta + \mu_{l(x_{ij})} \right],$$

where, as previously, α_i represents the tuple specific intercept, β the covariate effects and $\mu_{l(x_{ij})}$ the residual spatial 'height' associated with locations x_{ij} assigned to region l of the spatial partition. Recall, for a given partition, the function yielding the spatial heights is piecewise-constant taking the value μ_l in region $l \in 1, \dots, k$.

Now define

$$\lambda_i = \sum_{j=1}^m \lambda_{ij} = e^{\alpha_i} \sum_{j=1}^m e^{z'_{ij}\beta + \mu_{l(x_{ij})}}$$

and

$$y_i = \sum_{j=1}^m y_{ij}.$$

From standard distribution theory we know that the formulation in 3 is equivalent to

$$y_i \sim \text{Poisson}(\lambda_i)$$

and

$$(y_{i1}, \dots, y_{im} \mid y_i) \sim \text{Multinomial} \left(y_i, \frac{\lambda_{i1}}{\lambda_i}, \dots, \frac{\lambda_{im}}{\lambda_i} \right)$$

independently for $i = 1, \dots, n$. The conditioning thus yields a likelihood function which corresponds, up to proportionality in $\lambda_1, \dots, \lambda_n$, to the matched (conditional) likelihood in 2. More specifically, our data correspond to $y_{i1} = y_i = 1$ for all i and, hence, the likelihood becomes:

$$\prod_{i=1}^n e^{-\lambda_i} \lambda_i \left(1 + \sum_{j=2}^m \exp[(z_{ij} - z_{i1})' \beta + (\mu_{l(x_{ij})} - \mu_{l(x_{i1})})] \right)^{-1}.$$

To exploit this likelihood formulation a prior parameter specification which *a priori* maintains the product structure between $(\lambda_1, \dots, \lambda_n)$ and (β, μ) is necessary: $P(\beta, \mu, \lambda_1, \dots, \lambda_n) = P(\beta, \mu)P(\lambda_1, \dots, \lambda_n)$.

Now in terms of actual parameterisations both (β, μ, α) and (β, μ, λ) are equally workable, however, for the MCMC sampler it is easier to work in terms of the original parameterisation namely $(\beta, \mu, \alpha_1, \dots, \alpha_n)$. A priori taking

$$P(\lambda_i) \propto \frac{1}{\lambda_i}, i = 1, \dots, n,$$

leads to a joint prior distribution $P(\beta, \mu, \alpha_1, \dots, \alpha_n) \propto P(\beta, \mu)$, that is a flat prior on $(\alpha_1, \dots, \alpha_n)$. Hence, we can work with the model

$$y_{ij} \sim \text{Poisson} \left(\lambda_{ij} = e^{\alpha_i + z_{ij}' \beta + \mu_{l(x_{ij})}} \right)$$

with $i = 1, \dots, n, j = 1, \dots, m$, independent.

For $P(\beta, \mu)$ independent prior distributions are assumed and given by $p(\beta) \propto 1$ and $\mu_l \sim N(\mu_0, \sigma_0^2)$ where μ_0 is the empirical log-odds of disease, $\log \left(\frac{1}{m-1} \right)$. For σ_0^2 a hyper-prior is specified such that $\sigma_0^2 \sim IG(a, b)$ with $a = 1.0$ and $b = 0.01$ chosen on the basis of simulations.

3 Outline of MCMC Scheme

For the partition model, the dimensionality of the problem is unknown and thus a prior distribution for the number tiles needs to be specified. *A priori* the number of tiles, $k = 1, \dots, k_{max}$, assumes a truncated geometric distribution such that $P(k) \propto (1-t)^k$ where $t = 0.05$ was chosen on the basis of simulations. Conditional upon k , the locations of the tile centres, c_l , are taken to be uniform on A such that $P(c_l \mid k) = \|A\|^{-1}$ independent

for $l = 1, \dots, k$. A possible alternative to this prior specification would be to adopt a Poisson point process on A .

The posterior distribution is non-standard and of unknown dimension and thus a numerical approach for posterior sampling, based upon Reversible Jump MCMC, was implemented. Given the variable dimensionality of the model, reversible jump (Green, 1995) methodology was utilised. An extension of Gamerman's algorithm (1996) was developed to provide an efficient means of proposal. The methodology extends the methodology developed for handling the completely randomised case-control design.

As a means of exploring the joint posterior space the following moves were implemented:

Height: changes to the heights, $\mu_l, l = 1, \dots, k$, are proposed in succession.,

Birth: a unitary increase in the number of tiles, from k to $k + 1$, is proposed by means of sampling an additional tile centre c_{k+1} .,

Death: a unitary decrease in the number of tiles is proposed by means of deleting a single tile centre at random.,

Shift: tile centres $c_l, l = 1, \dots, k$, are proposed to be relocated locally within a sub-regional square.,

Beta: a proposal to update the β values *en bloc* is made.,

Alpha: nuisance, tuple-specific, parameters are updated in succession.,

Hyper: values of the hyper-parameters a, b are updated.

A random scanning mechanism is utilised for all but the *hyper* move, which is updated at each iterative step. At each stage of the MCMC algorithm non-Hyper move types are proposed with specific probability

$$\pi_{move}(k) : \sum \pi_{move}(k) = 1$$

the summation being over all implemented move types and the dependence upon k resulting from the incapacity for adding an additional tile centre when $k = k_{max}$ and, conversely, a death when $k = 1$.

The proposed new state was accepted according to the Metropolis-Hastings-Green ratio (Green, 1995) with the exception of the tuple specific parameters, $\alpha_i, i = 1, \dots, n$, and the hyper-parameters, $a, b : \sigma_0^2 \sim IG(a, b)$ which were updated by means of a Gibbs step. Computations of acceptance probabilities were performed on the log-scale, rounded and exponentiated to increase stability (Green, 1995).

The sampler was run for a total of 1,010,000 iterations of which the first 10,000 constituted the 'burn-in' period. Thinning was used to reduce dependency; with every 1000th sample generated being stored.

Simulation studies were conducted and indicated the methodological capacity to recover both smooth and discontinuity in the underlying risk surface.

4 Application: spatial variation in infant mortality risk, adjusted for Carstairs index

Infants are arguably the most vulnerable group in terms of the adverse effects of environmental health and thus the infant mortality rate has long been considered a general indicator of the level of development and quality of life of a given population. An increase in infant mortality rate may be taken to imply deterioration in that environment.

Infant mortality is usually expressed as the death rate of infants under one year of age per 1000 live births for a given time interval. Typically infants are described in terms of neonatal (under 28 days of age) and post-neonatal (aged 28 days to 1-year) deaths. Neonatal deaths can be further categorised as perinatal (within 7 days).

In this section the partition model methodology is applied to a dataset on perinatal mortality arising in the North-West of England. The data derive from the period 1980 to 1981 and consist of 3425 cases $y = 1$ of still births or deaths with seven days of life together with 6850 ($m=2$) live-birth controls $y = 0$ who were individually matched according to gender and date of birth. Spatial location x was geo-referenced according to the mothers place of residence, at the time of birth. Additional covariate information z consisted of the Carstairs score (Carstairs and Morris, 1991) devised to provide a measure of socio-economic deprivation. This score is based upon the percentage of persons, according to the 1991 census having no car, living in over-crowded housing and having a head of household in social class IV, V or unemployed. Low scores indicate increased affluence, whereas, high scores indicate increased deprivation. A point map, Figure 1, depicting a random sample of 400 controls and their matched controls is provided.

4.1 Results

The estimated odds ratio surface estimate for the perinatal data derived for the North West Thames Valley region is presented in Figure 2. The risk surface estimates were based upon the 'normalised' posterior mean odds, relative to the surface mean, at pixels on a 40×40 grid. The posterior distribution, with traces, for the effect of Carstairs Index on perinatal mortality provided in Figure 3

When adjusted for Carstairs index the residual posterior mean odds ratio, Figure 2, surface did not yield evidence of spatial variation in the odds of being still born or dying in the perinatal period in the North West Thames region. The odds ratio contours spanned 0.996, in the South, to 1.004, in the North West.

The posterior distribution for the effect of Carstairs Index, with traces, is presented in Figure 3. An increase in Carstairs index was found to be significantly associated with a reduction in perinatal mortality risk. The

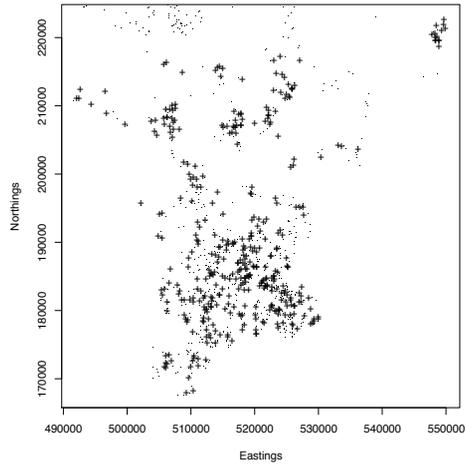


FIGURE 1. Spatial locations of 400 infants sampled using simple randomisation from those 3425 dying in the perinatal period together with their 800 matched controls. Cases are denoted with a '+' and cases by a '.'

posterior mean $\tilde{\beta}$ was found to be 0.047, with a posterior standard deviation of 0.0072, corresponding to a reduction in the odds of death in this vulnerable period of life of approximately 5% for a unit increase in Carstairs score.

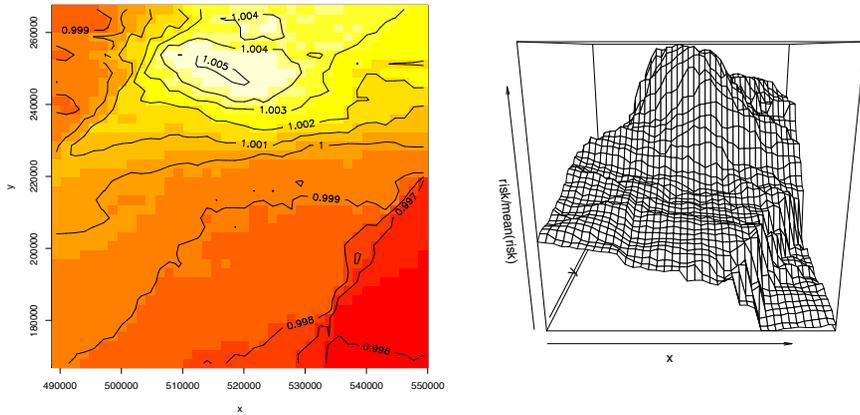


FIGURE 2. Results of analysis based upon infant mortality data, adopting a truncated geometric prior for k with t fixed at 0.1 and the dispersion hyper-parameters fixed such that $\sigma_0^2 \sim (1, 0.01)$.

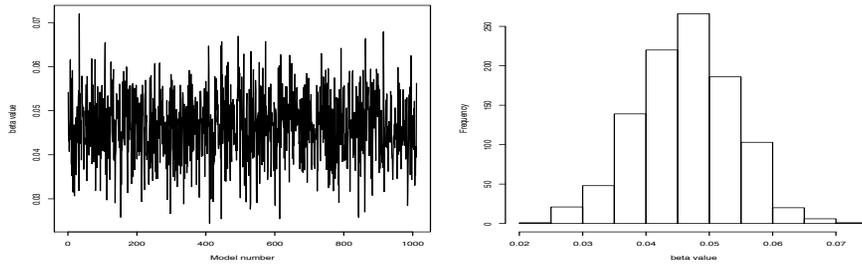


FIGURE 3. Posterior distribution with traces for the effect of Carstairs index on mortality risk based upon a truncated geometric prior for k with t fixed at 0.1 and the dispersion hyper-parameters fixed such that $\sigma_0^2 \sim (1, 0.01)$ was used.

Acknowledgments: Special thanks to Leonhard Held and Carmen Fernandez for their support prior to and during this work

References

- Breslow, N.E. and Day, N.E. (1980). Statistical methods in cancer research. Vol 1. The analysis of case-control studies IARC *Scientific publications* No. 32.
- Costain, D.A. (accepted September 2008) Bayesian partitioning for modelling and mapping spatial case-control data. *Biometrics* to appear (available in from early view)
- Gamerman, D. (1996). Sampling from the posterior distribution in generalised linear mixed models. *Statistics and Computing* 7, 57-68.
- Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82, 4, pp. 711-731.

A Bayesian Approach to Analysis of Covariance for Split-Plot Designs

Caitlin M. Cunningham¹ and James G. Booth¹

¹ Department of Statistical Sciences, 301 Malott Hall, Cornell University, Ithaca, NY 14850, Phone: (781)534-2324, Fax: (607)255-9801, email: cc487@cornell.edu

1 Introduction

Analysis of covariance in designed experiments has a long history dating back to Fisher. Given the popularity of Bayesian approaches to statistical modelling and inference, it is somewhat surprising that there is so little literature on the application of Bayesian methods in this context. A recent paper by Wang and Hsu (2006) develops Monte Carlo methods for Bayesian analysis of randomized complete block designs. However, the extension to settings in which there is covariate information is not straightforward. In Cunningham and Booth (2009), we develop such an extension by building upon a multivariate mixed model for designed experiments proposed by Booth et al (2009). While the methodology was generalized to include the balanced incomplete block design setting, we now consider a more complex experimental design for split-plots.

2 Development

Consider a split-plot design with response variable, Y , and a measured covariate, Z . Suppose that Z was measured prior to the application of the treatment, but that it may be correlated with the response. An “obvious” linear model (Milliken and Johnson, 1984, p.397) for a split-plot experiment with a whole-plot treatments, t split-plot treatments, and b whole-plots assigned to each whole-plot treatment group is:

$$Y_{ij} = \mu + \alpha_i + \tau_k + B_{(i)j} + \alpha\tau_{ik} + \gamma z_{ijk} + E_{ijk} \quad (1)$$

where μ is the overall mean, $\alpha_i, i = 1, \dots, a$ are the whole-plot treatment effects, $\tau_k, k = 1, \dots, t$, are the split-plot treatment effects, $\alpha\tau_{ik}$ are the treatment interaction effects, and γ is the effect of regression on the covariate Z . $B_{(i)j}, j = 1, \dots, b$, are the random whole-plot effects nested within whole-plot treatment and assumed to be independent and identically distributed (i.i.d.) normal variables with mean zero and variance σ_b^2 and E_{ijk}

are the random errors assumed to be i.i.d. normal variables with mean zero and variance σ_e^2 . In addition, we impose the standard identifiability (sum) constraints: $\sum_i \alpha_i = 0$, $\sum_k \tau_k = 0$, $\sum_i \alpha\tau_{ik} = 0$ and $\sum_k \alpha\tau_{ik} = 0$.

Note that model 1 is inherently conditional on the values of the covariate Z . This begs the question: what joint distribution for Y and Z gives rise to these models? This question is answered by Booth et al. (2009) who consider a bivariate model in which the distribution of Z is independent of the treatments, but which allows for random variation between the whole-plots and residual error; specifically

$$\begin{pmatrix} Y_{ijk} \\ Z_{ijk} \end{pmatrix} = \begin{pmatrix} \mu_y \\ \mu_z \end{pmatrix} + \begin{pmatrix} \alpha_{i,y} \\ 0 \end{pmatrix} + \begin{pmatrix} \tau_{k,y} \\ 0 \end{pmatrix} + \begin{pmatrix} \alpha\tau_{ik,y} \\ 0 \end{pmatrix} \quad (2) \\ + \begin{pmatrix} B_{(i)j,y} \\ B_{(i)j,z} \end{pmatrix} + \begin{pmatrix} E_{ijk,y} \\ E_{ijk,z} \end{pmatrix}$$

where the whole-plot effects are i.i.d bivariate normal,

$$\begin{pmatrix} B_{(i)j,y} \\ B_{(i)j,z} \end{pmatrix} \sim \text{i.i.d. } N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_b = \begin{pmatrix} \sigma_{b,y}^2 & \sigma_{b,yz} \\ \sigma_{b,yz} & \sigma_{b,z}^2 \end{pmatrix} \right]$$

independently of the bivariate residual errors

$$\begin{pmatrix} E_{ijk,y} \\ E_{ijk,z} \end{pmatrix} \sim \text{i.i.d. } N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_e = \begin{pmatrix} \sigma_{e,y}^2 & \sigma_{e,yz} \\ \sigma_{e,yz} & \sigma_{e,z}^2 \end{pmatrix} \right].$$

This joint model implies a univariate conditional model of the form

$$Y_{ij} = \mu + \alpha_i + \tau_k + \alpha\tau_{ik} + B_{(i)j} + \gamma_e Z_{ijk} + \gamma_b \bar{Z}_{ij} + E_{ijk}, \quad (3)$$

where $\gamma_e = \sigma_{e,yz}/\sigma_{e,z}^2$ and $\gamma_e + \gamma_b = (\sigma_{b,yz} + \sigma_{e,yz}/t)/(\sigma_{b,z}^2 + \sigma_{e,z}^2/t)$ are respectively the slopes for intra- and inter-whole-plot regressions. Model 1 is the special case in which $\sigma_{b,z}^2 = 0$ (and hence also $\sigma_{b,yz} = 0$). Thus, model 1 implicitly assumes that there is no between whole-plot variation in the covariate means, a very unrealistic situation in practice. Booth et al. (2009) show that estimation using model 1 results in (conditionally) biased adjusted treatment means and incorrect standard errors. In contrast, estimation using the bivariate model, or the implied univariate conditional model 3, produces unbiased adjusted means and appropriate standard errors.

As Wang and Hsu (2006) discuss in their paper on Bayesian ANOVA, in the case when some prior information about the experiment is known, a Bayesian analysis may be desired. In Cunningham and Booth (2009), we explore a Bayesian analysis of the randomized complete block design and incomplete block designs, and use two simple algorithms and Gibbs sampling to sample from the posterior distributions. We follow the development in that paper to find the posterior distributions for the split-plot design case. By using conjugate priors, we find that the posterior distributions of the

whole-plot and split-plot treatment effects are both multivariate-t (subject to a constraint). Both of these distributions are easy to sample from, thereby allowing quick estimation of posterior fixed effects.

3 Likelihood

We can divide the data into three independent parts:

$$\begin{aligned} \text{treatment} &: \begin{pmatrix} \{\bar{Y}_{i \cdot k}\} \\ \{\bar{Z}_{i \cdot k}\} \end{pmatrix} \\ \text{whole-plot} &: \begin{pmatrix} \{\bar{Y}_{ij \cdot} - \bar{Y}_{i \cdot \cdot}\} \\ \{\bar{Z}_{ij \cdot} - \bar{Z}_{i \cdot \cdot}\} \end{pmatrix} \\ \text{error} &: \begin{pmatrix} \{Y_{ijk} - \bar{Y}_{i \cdot k} - \bar{Y}_{ij \cdot} - \bar{Y}_{i \cdot \cdot}\} \\ \{Z_{ijk} - \bar{Z}_{i \cdot k} - \bar{Z}_{ij \cdot} - \bar{Z}_{i \cdot \cdot}\} \end{pmatrix}. \end{aligned}$$

Furthermore, we can show these three matrices are distributed as follows:

$$\begin{aligned} \text{treatment} &\sim N(\theta, (\frac{1}{b}(\Sigma_e \otimes I_t + \Sigma_b \otimes J_t) \otimes I_a)) \\ \text{SSW} = b \times (\text{whole-plot}) \times t(\text{whole-plot}) &\sim W_2(a(b-1), \Sigma_{be}) \\ \text{SSE} = \text{error} \times t(\text{error}) &\sim W_2(a(b-1)(t-1), \Sigma_e) \end{aligned}$$

where $\theta_{ik,y}$ is $\mu_y + \alpha_{i,y} + \tau_{k,y} + \alpha\tau_{ik,y}$, $\Sigma_{be} = \Sigma_e + t\Sigma_b$, J_t is the t by t matrix of ones, and $W_p(v, \Sigma)$ is the Wishart distribution for a p by p matrix with degrees of freedom v and scale matrix Σ .

Let $\gamma_{ik,y} = \tau_{i,y} + \alpha\tau_{ik,y}$. For simplicity's sake, we will first assume that τ_z , α_z and $\alpha\tau_z$ are non-zero unknown parameters. We will set them to zero later in the development. The likelihood can be written as the product of the three independent distributions of the data above, and can be rewritten so that it consists of two independent pieces - one for $(\bar{\theta}, \Sigma_{be})$ and one for (γ, Σ_e) .

4 Priors and Posteriors

Following Wang & Hsu (2006), the following conjugate priors are selected:

- (i) Given Σ_{be} with hyper-parameters $\mu_{\bar{\theta}}$ and c , $\bar{\theta} \sim N_{2a}(\mu_{\bar{\theta}}, c\Sigma_{be} \otimes I_a)$.
- (ii) Given Σ_e with hyper-parameters μ_γ and C , $\gamma \sim N_{2(t-1)}(\mu_\gamma, \Sigma_{be} \otimes C)$.
- (iii) Given Σ_e and hyper-parameters ν_1 and Λ_1 , $\Sigma_{be} \sim \text{inv} - \text{Wish}_2(\Lambda_1^{-1}, \nu_1)$, where $\Sigma_{be} - \Sigma_e$ is positive definite.
- (iv) With hyper-parameters ν_2 and Λ_2 , $\Sigma_e \sim \text{inv} - \text{Wish}_2(\Lambda_2^{-1}, \nu_2)$.

The full joint posterior is of the same form of the prior, with the hyper-parameters updated.

We can divide the full joint posterior into two independent functions, one that depends on $(\bar{\theta}, \Sigma_{be})$ and one that depends on (γ, Σ_e) . We are interested in the marginal posterior for both γ and $\bar{\theta}$. With these two quantities, all the mean parameters can be estimated.

We find that both $(\bar{\theta}_y|\bar{\theta}_z)$ and $(\bar{\theta}_z|\bar{\theta}_y)$ are multivariate-t's, and so can use a Gibbs Sampler to sample from each of them. We also note that $\theta_{i,z}$ is only equal to μ_z , as we know that $\alpha_{i,z}$ is zero, and take this into account. For γ , we note that γ_z is known to be zero. The marginal posterior distribution of $(\gamma_y|\gamma_z = 0)$ is also multivariate-t.

5 Letting Treatment Means be Zero

In the above development, priors were placed on the values γ_z , τ_z , and $\alpha\tau_z$, although it is known that these values are in fact zero. Setting them to zero at the outset forces a new selection of priors and creates a slower algorithm; however, it is also more accurate. In particular, the priors for $\bar{\theta}$ and γ_y can no longer depend on Σ_e and Σ_{be} . Since $\bar{\theta}_z$ is equal only to μ_z , it is no longer a vector. The selected priors are as follows:

- (1) Given hyper-parameters $\mu_{\bar{\theta}}$ and $C_{\bar{\theta}}$, $\bar{\theta} \sim N_{a+1} \left(\mu_{\bar{\theta}}, C_{\bar{\theta}} = \begin{pmatrix} C_{\bar{\theta},y} & c_{\bar{\theta},zy} \\ c_{\bar{\theta},yz}^T & c_{\bar{\theta},z} \end{pmatrix} \right)$.
- (2) Given hyper-parameters $\mu_{\gamma,y}$ and C_{γ} , $\gamma_y \sim N_{t-1}(\mu_{\gamma,y}, C_{\gamma})$.

The priors for Σ_{be} and Σ_e remain the same.

Once again, it is necessary to find the three posterior distributions $(\bar{\theta}_y|\bar{\theta}_z)$, $(\bar{\theta}_z|\bar{\theta}_y)$ and (γ_y) . They each are of the form normal times multivariate-t, which means they cannot be directly sampled from. Instead, we use an accept reject algorithm to sample from these posteriors. It is because of this that the algorithm for this method (method 2) can take much longer than for method 1 (in which the treatment means were not set to zero at the outset).

6 Programming and Results

We develop two algorithms, one for each of the methods described above.

6.1 Method 1

Recall that the above distributions were found ignoring the constraint that $\Sigma_{be} - \Sigma_e$ must be positive definite. We can generate posterior treatment means taking this constraint into account using the following algorithm:

- (1) Select a starting value for $\bar{\theta}_z$.
- (2) Generate a proposed γ_y from the posterior for $(\gamma_y|\gamma_z = 0)$.
- (3a) Generate a proposed $\bar{\theta}_y$ from the conditional posterior $(\bar{\theta}_y|\bar{\theta}_z)$.
- (3b) Generate a proposed $\bar{\theta}_z$ from the conditional posterior $(\bar{\theta}_z|\bar{\theta}_y)$.
- (4a) Generate a proposed Σ_e from the conditional posterior $(\Sigma_e|\bar{\theta}, \gamma)$.
- (4b) Generate a proposed Σ_{be} from the conditional posterior $(\Sigma_{be}|\bar{\theta}, \gamma)$.
- (5) Check if the proposed $\Sigma_e - \Sigma_{be}$ is positive definite. If so, save the proposed $(\bar{\theta}, \gamma)$, else, do not save.
- (6) If number of saved $(\bar{\theta}, \gamma)$ less than desired sample, return to 2. Else, end.

Note that we need the conditional posterior distributions of Σ_{be} and Σ_e for this algorithm. They can easily be found, and are both inverse-Wisharts.

6.2 Method 2

For Method 2, we need three accept-reject algorithms to generate from the needed distributions. In each case, we use the multivariate-t part of the posterior distributions as the proposal distribution, and then keep the proposal if $r < U$, where U is a random uniform on $(0, 1)$ and r is the exponential part of the normal distribution.

- (1) Select a starting value for $\bar{\theta}_z$.
- (2) Generate a proposed γ_y from the posterior for γ_y using an accept reject algorithm.
- (3a) Generate a proposed $\bar{\theta}_y$ from the conditional posterior $(\bar{\theta}_y|\bar{\theta}_z)$ using an accept reject algorithm.
- (3b) Generate a proposed $\bar{\theta}_z$ from the conditional posterior $(\bar{\theta}_z|\bar{\theta}_y)$ using an accept reject algorithm.
- (4a) Generate a proposed Σ_e from the conditional posterior $(\Sigma_e|\bar{\theta}, \gamma_y)$.
- (4b) Generate a proposed Σ_{be} from the conditional posterior $(\Sigma_{be}|\bar{\theta}, \gamma_y)$.
- (5) Check if the proposed $\Sigma_e - \Sigma_{be}$ is positive definite. If so, save the proposed $(\bar{\theta}, \gamma)$, else, do not save.
- (6) If number of saved $(\bar{\theta}, \gamma)$ less than desired sample, return to 2. Else, end.

Once again, the conditional distributions for Σ_e and Σ_{be} are needed. They are found to be inverse-Wisharts.

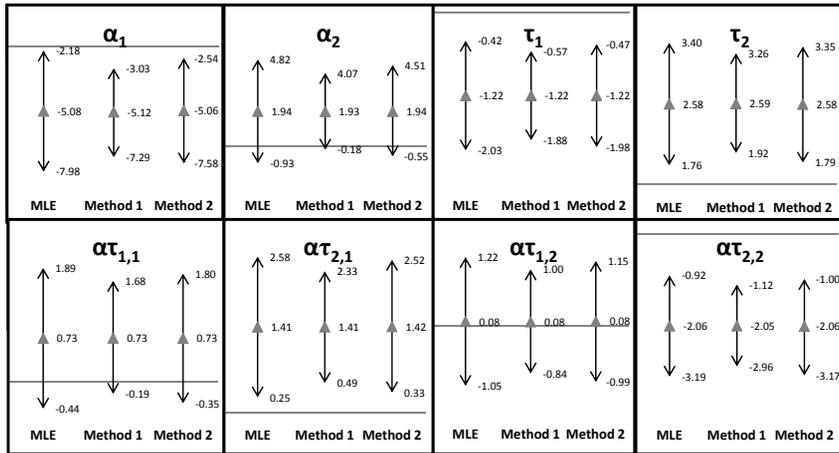


FIGURE 1. Cookie Data Parameters

7 Data

We consider a split-plot experimental design presented in Milliken & Johnson (2002) involving cookie baking. Four ovens each bake cookies of three different types at three different temperatures. The diameter of the cookies are measured after baking as a response, and the thickness of each cookie before baking is measured as a possible covariate. In this case, the whole-plot treatment is the temperature, the whole-plot is the oven, and the treatments are the different cookie types. We run both our model and use the statistical software JMP to estimate the model parameter values and to compare the Bayesian estimates and the MLEs. Vague priors are used, so the Bayesian estimates will be close to the MLEs.

Figure 1 shows the parameter estimates, excluding those that can be found using the sum constraints. In this case, α is the treatment effect of temperature, τ is the treatment effect of cookie type, and $\alpha\tau$ is the interaction between them. The figures below show the parameter estimates for each, along with the upper and lower confidence/credible bounds. The horizontal line on each panel represents zero - when it crosses the confidence intervals, we can conclude that those parameters are not significant. Note that the credible interval for Method 1 is in general shorter than that for the MLE and for Method 2. This is due to the unnecessary priors placed on τ_z and α_z in method 1, which causes the multivariate-t to have extra degrees of freedom, thus underestimating the standard error. As a , the number of whole-plot treatments, or b , the number of whole-plots, gets large, this difference becomes negligible. With a large data set with a large number of whole-plots, method 1 and method 2 produce very similar results. This is important, because method 2 can take longer to run, due to the additional

monte carlo steps needed for that algorithm.

Each distribution was run for 10,000 iterations, so the Monte Carlo standard error can be estimated by $SD/100$. The cookie data set took about 20 minutes to run for each method because a large number of samples had to be discarded due to a violation of the constraint that $\Sigma_{be} - \Sigma_e$ must be positive definite. To produce 10,000 samples, the algorithm for the cookie data set had to propose 37,250 iterations. This algorithm was written in the statistical computing language R; using a different language, such as C, would significantly speed up computing time as well.

References

- Booth, James G. and Federer, Walter T. and Wells, Martin T. and Wolfinger, Russell D. (2009). A Multivariate Variance Components Model for Analysis of Covariance in Designed Experiments. Under Review by *Statistical Science*.
- Cunningham, Caitlin M. and Booth, James G. (2009). A Bayesian Approach to Analysis of Covariance in Balanced Randomized Block Experiments. preprint.
- JMP, Version 7 . SAS Institute Inc., Cary, NC, 1989-2007.
- Milliken, George A. and Johnson, Dallas E. (1984). *Analysis of Messy Data Volume III: Analysis of Covariance*. New York: Chapman & Hall.
- Wang, Jenting and Hsu, John S. J. (1995). Bayesian Analysis of the Additive Mixed Model for Randomized Block Designs. *Australia and New Zealand Journal of Statistics*, **48**, 225-236.

A latent structure model for high river flows

T. Economou¹, R. Vitolo¹, T. C. Bailey¹, E. Waterhouse²
and Z. Kapelan¹

¹ School of Engineering, Computer Science and Mathematics, University of Exeter, Harrison Building, North Park Road, Exeter, EX4 4QF, UK

² Department of Geography, Durham University, UK

Abstract: River discharge in the UK exhibit significant clustering of high-flow events on multidecadal timescales. A hidden semi-Markov model is constructed for the study of such multidecadal variability. The model includes time dependent covariates of climatological nature as well as a random effect driven by the hidden Markov states to account for possible non-explicit low-frequency climatic processes. The model is applied to an illustrative data set for river Severn in the UK.

Keywords: Hidden semi-Markov models; floods; ultralow-frequency variability.

1 Introduction

Floods are natural catastrophes which may have a disastrous effect in economic terms. For example, UK floods in summer 2007 resulted in the largest flood-related aggregate insured loss in the UK (Lane, 2007). These events have generated considerable concern in the insurance industry and, therefore, interest in better understanding the associated statistical properties. The problem of determining return periods for high-flow river discharge is particularly delicate if the underlying physical process is intrinsically non-stationary, for example due to climate change or to anthropogenic causes (e.g. change in land usage). A recent study of a number of catchments in UK has revealed remarkable hydrological volatility in the past: major floods appear to be characterised by significant spatio-temporal clustering (Robson, 2002). The pronounced temporal variability of flood occurrences has been described in terms of ‘flood rich’ and ‘flood poor’ periods which may extend for multiple decades (Robson, 2002; Lane, 2007).

Low-frequency behaviour at decadal timescales is a possible explanation for the clustering behaviour described above, somewhat contradicting the climate change hypothesis. Recent increasing trends in the frequency of flooding in certain catchments may be explained as irregularly recurring patterns of variability, occurring on multidecadal timescales. Climatic variability at such low frequencies has indeed been observed and described in

a fairly large number of studies and it is crucial to understand which climatic trends can be attributed to natural variability of the climate system, rather than to anthropogenic forcing. In this paper we consider a hidden semi-Markov model in an effort to try and capture features of the variability in flooding that may be driven by unobserved processes.

2 Model Specification

Consider a data set which is a time series $\mathbf{y} = \{y_1, y_2, \dots, y_N\}$ of yearly counts of single days for which a flood event was recorded assuming that the river was observed for N years. Here we assume a non-stationary Poisson model for y_t ($t = 1, \dots, N$), where the mean $\Lambda(x_t)$ may depend on possibly time dependent covariates x_t . Furthermore we assume that the mean depends on a hidden state S_t of a semi-Markov chain at time t where $S_t \in \{1, 2, \dots, S\}$ is the state space of the process. The mean is characterised as follows:

$$\Lambda(\mathbf{x}_t; S_t) = \exp\{\theta_{S_t} + \beta \mathbf{x}_t\}$$

so that given the state S_t , y_t is Poisson distributed with a mean that depends on time through covariates $\mathbf{x}_t = (x_{1t}, x_{2t}, \dots, x_{pt})$ which have associated parameters $\beta = (\beta_1, \dots, \beta_p)$. According to the state of a hidden semi-Markov chain, $\Lambda(\mathbf{x}_t; S_t)$ will be different for each state through the state dependent intercept θ_{S_t} . Here we assume that the resulting hidden semi-Markov Poisson (HSMP) model occurs in discrete time mainly due to the nature of flood data but also because it reduces the complexity of the model in a way. Note that through the hidden chain, some correlation structure is introduced in the counts y_t .

To derive the likelihood of the HSMP model consider first the likelihood of the Poisson model over a period τ given the state s of the chain during that period:

$$\ell(y_1, y_2, \dots, y_\tau | s) = \prod_{i=1}^{\tau} \frac{e^{-\Lambda(\mathbf{x}_i; s)} \Lambda(\mathbf{x}_i; s)^{y_i}}{y_i!} \quad (1)$$

Second, consider a simple semi-Markov chain which is often defined by an initial state distribution $\boldsymbol{\pi} = (\pi(1), \pi(2), \dots, \pi(S))$, a transition probability matrix $\mathbf{P} = \{p_{ij}\}$ where $p_{ii} = 0$, $\sum_j p_{ij} = 1$ and a vector of holding time distributions $\mathbf{h}(\tau) = \{h_i(\tau)\}$. So the chain starts at a state s_1 say, according to $\pi(s_1)$ and holds that state for a time interval τ_{s_1} according to distribution $h_{s_1}(\tau_{s_1})$, it then enters a new state s_2 according p_{s_1, s_2} and the process repeats itself analogously. The likelihood of a realisation $(\tau_{s_1}, \tau_{s_2}, \dots, \tau_{s_n})$ of this chain involving n state changes is

$$\pi(s_1) h_{s_1}(\tau_{s_1}) \prod_{j=2}^n p_{s_{j-1}, s_j} h_{s_j}(\tau_{s_j}) \quad (2)$$

Note that $h_i(\tau)$ can be any discrete distribution and if it is geometric then the chain is Markov and not semi-Markov. The reason for considering a semi-Markov chain here is because it increases model flexibility by not imposing a specific structure on $h_i(\tau)$.

Now suppose that both the time series \mathbf{y} and the semi-Markov chain have been observed. Then the joint likelihood $L(y_1, \dots, y_N; \tau_{s_1}, \dots, \tau_{s_n})$ is obtained by combining (1) and (2):

$$\pi(s_1)h_{s_1}(\tau_{s_1})\ell(y_1, \dots, y_{\tau_{s_1}} | s_1)p_{s_1, s_2}h_{s_2}(\tau_{s_2})\ell(y_{\tau_{s_1}+1}, \dots, y_{\tau_{s_1}+\tau_{s_2}} | s_2) \times \text{etc.}$$

But since the chain is not observed, we sum over all possible states $s_i \in \{1, 2, \dots, S\}$ and all possible holding times $\tau_i \in \{1, 2, \dots, \infty\}$ to obtain the marginal likelihood $L(y_1, \dots, y_N)$ of the HSMP model as:

$$\sum_{\tau_{s_1}=1}^{\infty} \dots \sum_{\tau_{s_n}=1}^{\infty} \sum_{s_1=1}^S \dots \sum_{s_n=1}^S L(y_1, \dots, y_N; \tau_{s_1}, \dots, \tau_{s_n}) \quad (3)$$

In general, the HSMP likelihood is a function of the parameters in $\Lambda(\mathbf{x}(t); S_t)$, the unknown initial distribution $\boldsymbol{\pi}$ and transition matrix \mathbf{P} and also the parameters $\boldsymbol{\phi}$ of the specified holding distributions $h_i(\tau_i)$. Given the complexity of the model we adopt an MCMC approach to model fitting considering that the computational cost in evaluating (3) is great given any reasonable observation interval and number of proposed states. By our assumption, the data \mathbf{y} for the HSMP model are expressed in discrete time steps meaning that recursive algorithms used in the Hidden Markov models literature (MacDonald and Zucchini, 1997) can be analogously modified for efficient calculation of the HSMP likelihood. The idea in such recursion is to consider a variable $\alpha_t(j)$ sequentially at each discrete time step $t \in \{1, \dots, N\}$, where:

$$\alpha_t(j) = \Pr(y_1, \dots, y_t \text{ and chain exits state } j \text{ at time } t)$$

One can then compute $\alpha_t(j)$ recursively:

$$\begin{aligned} \alpha_1(j) &= \pi(j)h_j(1)\ell(y_1|j) \\ \alpha_2(j) &= \pi(j)h_j(2)\ell(y_1, y_2|j) + \sum_{i \neq j} \alpha_1(i)p_{ij}h_j(1)\ell(y_2|j) \\ \alpha_3(j) &= \text{etc...} \end{aligned}$$

Then $\sum_{j=1}^S \alpha_N(j)$ is equivalent to (3). Note that in the recursive expression for $\alpha_N(j)$, we replace $h_j(\cdot)$ with its upper tail to account for right censoring in the final state duration. Once the likelihood is efficiently evaluated, it can be used in conjunction with Metropolis-Hastings to provide a computationally feasible estimation procedure for the parameters of the HSMP model. Here we use a combination of the random walk and the independence Metropolis samplers (Gilks et al., 1996) but do not provide any details.

3 Model Application

We examine daily discharge data for the river Severn at Bewdley (UK) in the 85 year period between 1922 and 2006. A flood is recorded when the discharge passes a certain threshold and the response \mathbf{y} is defined as the number of days a flood has occurred in a year. Covariates $x_1(t)$ and $x_2(t)$ are used corresponding to the yearly averages of Atlantic multidecadal oscillation (AMO) and North Atlantic oscillation (NAO) indexes between 1922 and 2006 respectively. The possible presence of other, not explicit low-frequency processes is accounted for by the hidden semi-Markov chain. Specifically we assume two hidden states S_t in the chain where each has a Poisson holding time with a different mean. The model is then

$$y_t \sim \text{Pois}(\Lambda(x_{1t}, x_{2t}; S_t)) \quad t = 1, \dots, 85$$

$$\Lambda(x_{1t}, x_{2t}; S_t) = \exp\{\theta_{S_t} + \beta_1 x_{1t} + \beta_2 x_{2t}\} \quad S_t \in \{1, 2\}$$

10000 samples were collected from the posterior distribution of each parameter and from the posterior predictive distributions of the fitted values. In figure (1) the black line represents the actual values \mathbf{y} , the dashed line shows the fitted values calculated as the means of the posteriors and bold lines show the 95% credible intervals calculated as the 95% quantiles of the posteriors (note that the lower interval is 0 for all years)

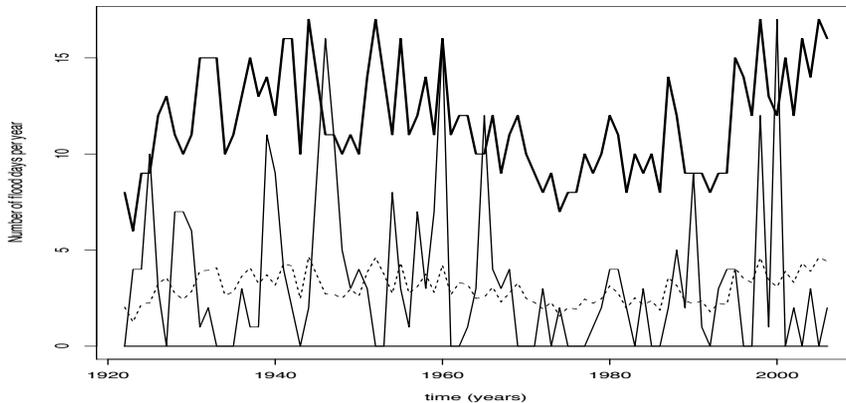


FIGURE 1. Fitted and actual values of the response.

3.1 Conclusion and Discussion

Figure 1 shows that the model is able to capture the increased variance in prevalence between 1930 and 1960 and in the last part of the record: these

are two ‘flood-rich’ periods for the Severn, separated by a long ‘flood-poor’ period between 1960 and the late 1990s. Although this behaviour is mainly explained by the covariate AMO and not the hidden chain, the model did identify two hidden states, one with a very small prevalence in time but with a higher value of θ_{S_t} in the Poisson mean which is what the data is suggesting looking at the ‘spikes’ of large values in the observed counts in Figure (1). This is reflected in the sufficiently high credible intervals. Possible extensions to the model include the introduction of a (spatial) random effect to facilitate a multi-catchment scenario. Also, one of the main point of interest will be to try and characterise the hidden states of the Markov model in terms of physical processes which would be possibly useful in setting up improved seasonal or interannual forecasts of high-flows.

Acknowledgments: The authors are indebted to Prof S. Lane for insightful discussions.

References

- Gilks, W.R., Richardson S. and Spiegelhalter D.J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall
- Lane, S.N. (2007). The 2007 UK summer floods: a scientific perspective www.willisresearchnetwork.com/Lists/Publications/DispForm.aspx?ID=6.
- MacDonald, I. and Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series*. Chapman and Hall.
- Robson, A.J. (2002). Evidence for trends in UK flooding. *Philosophical transactions of the Royal Society A*, **360**, 1327.

Deconvolution of Spike Trains Using an L_0 Penalty

Paul H. C. Eilers¹

¹ Department of Biostatistics, Erasmus Medical Centre, Rotterdam, The Netherlands, email: p.eilers@erasmusmc.nl

Abstract: Many chemical instruments deliver spike trains. Their heights and positions contain the information of interest. A useful model for spike trains is a convolution of a series of very sharp input pulses by a constant impulse response. To estimate the input, penalized regression with an L_0 norm is shown to work well. The impulse response can also be estimated from the data.

Keywords: Electrophoresis, Sequencing, Ill-conditioned, Blind deconvolution.

1 Introduction

Many instruments in chemical and biological laboratories produce signals that consist of a sum of spikes or pulses. The spikes have (more or less) equal shapes but different heights, and they may overlap. An example, from DNA sequencing, is shown in Figure 5. A useful model for such a signal is a series of idealized spikes (impulse functions) that have been fed through a linear system with a fixed impulse response. It is of interest to estimate positions and heights of the input spikes.

Assuming that the impulse response of the system is known, this looks like a simple case of regression, but it is not. For real data the problem is severely ill-conditioned and the results of straightforward regression are useless. Li and Speed (2004) discuss the condition of and present solutions. One possibility is to model the signal parametrically as a sum of shifted and scaled replicas of the impulse response. If the latter is known accurately, this is an effective approach. However, the model is highly non-linear in the parameters for the shifts and really good starting estimates are needed. Cardot et al. (2004) propose to use boosting in this context.

Li and Speed (2004) also propose regression under constraints and they discuss connections to the EM algorithm that is popular in image de-blurring (Vardi and Lee, 1993). This algorithm iteratively redistributes the observed output, proportionally to the current estimate of the input. Averaging gives an improved estimate of the input, to be used in the next iteration.

An alternative out is to use penalties. A ridge or L_2 penalty shrinks the size, as measured by the sums of squares, of the estimated input signal.

It removes the ill condition, but the result does not really look like a train of isolated narrow spikes. An L_1 norm, the sum of the absolute values, in penalty gives a big improvement, but still one generally does not get the desired result.

I like to draw attention to deconvolution using the L_0 norm, effectively penalizing the number of non-zero values in the estimated input. It works amazingly well, and it can be implemented by a simple iterative algorithm. It can not always be assumed that the impulse response is available. That leads to so-called blind deconvolution: both the input and the impulse response have to be estimated from the output signal. If we know the input signal, we can estimate the impulse response by (well-conditioned) regression. This suggests an iterative algorithm, alternating between estimating input and impulse response. This turns out to work well, if a reasonable starting estimate of the impulse response is used.

In the next section I introduce the model and show its performance on real data. In the Discussion I point to complications and extensions

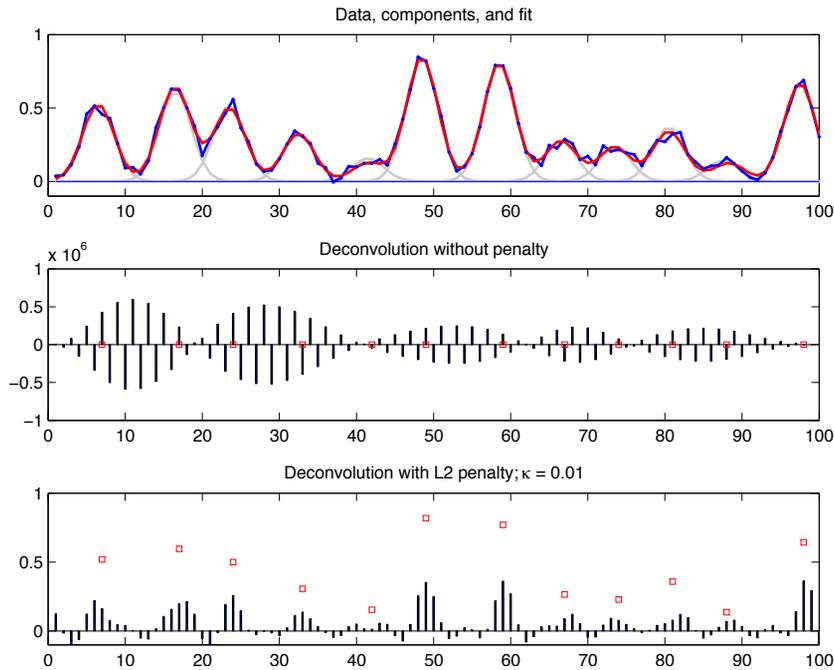


FIGURE 1. Simulated data. Top panel: data (full blue line with dots), fit (full red line) and individual pulses (thick gray lines). Middle panel: input as estimated without a penalty. Bottom panel: input as estimated without an L_2 penalty. The small squares give the positions and the heights of the non-zero elements of the input used for the simulation.

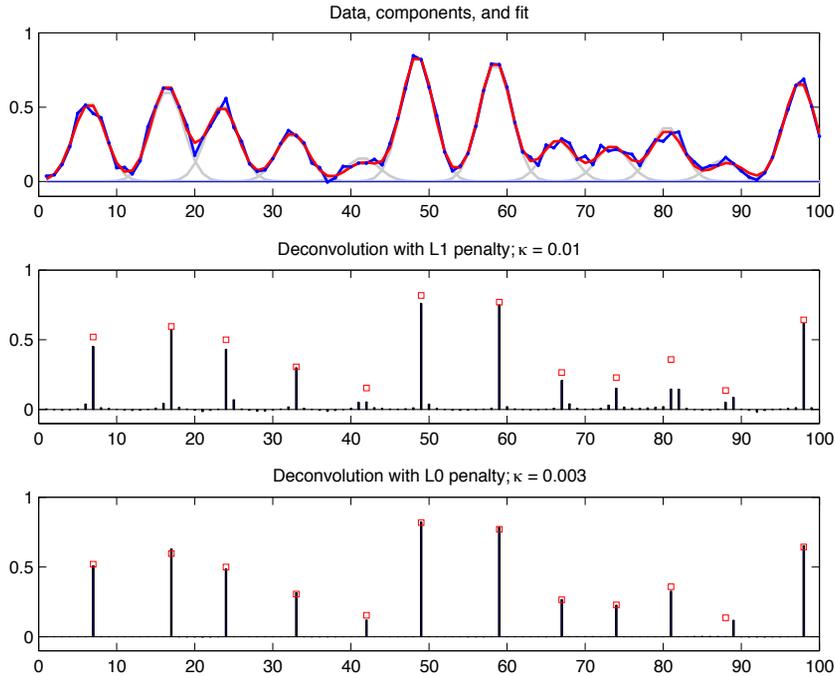


FIGURE 2. Simulated data. Top panel: data (full blue line with dots), fit (full red line) and individual pulses (thick gray lines). Middle panel: input as estimated with an L_1 penalty. Bottom panel: input as estimated with an L_0 penalty. The small squares give the positions and the heights of the non-zero elements of the input used for the simulation.

2 The model and its application

2.1 Convolution and deconvolution

Consider a (causal) discrete linear system with impulse responses c , and an input signal x . Then, by the superposition principle, the output signal y , of length m , is described by

$$y_i = \sum_{j=0}^n c_j x_{i-j}.$$

Details can be found in many books on linear system theory; see e.g. Brown (2007). Interpreting i as indexing time, this shows that y_i is a weighted sum of the present and all previous observations of x . The weights are given by the impulse response. In practice n is relatively small compared

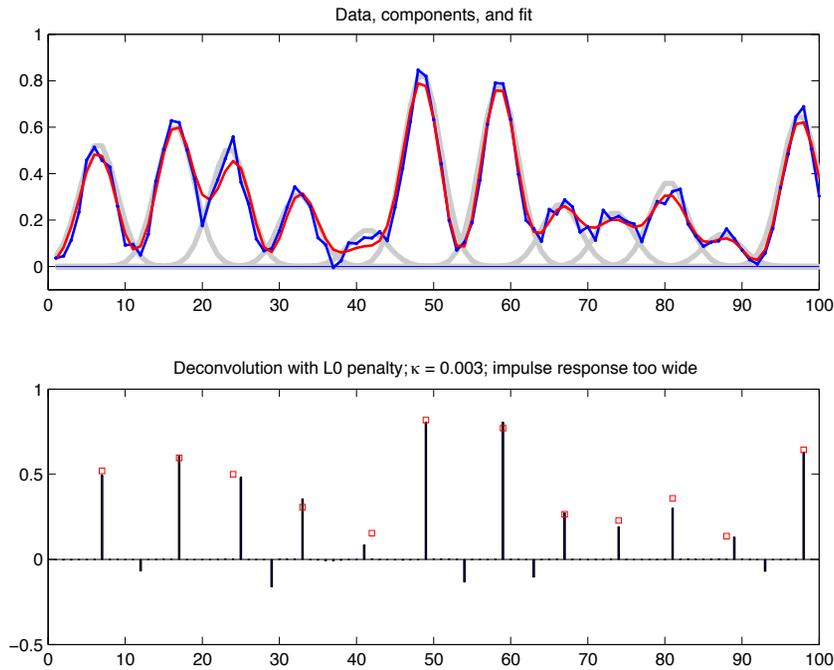


FIGURE 3. Simulated data. Upper panel: data (full blue line with dots), fit (full red line) and individual pulses (thick gray lines). Lower panel: deconvolution with an impulse response that is 20% too wide.

to the lengths of x and y . We can write $y = Sx$, where S is a matrix with m rows and $m + n$ columns with $s_{ij} = c_{n-j+1}$ if $n - j + 1 \leq 0$ and $s_{ij} = 0$ otherwise. Row i of S contains c reversed and right-shifted by $i - 1$. Consider the following inverse problem: y and c are given, find x . This looks like a simple regression problem. In fact it is, but it is extremely ill conditioned. Figure 1 illustrates this for simulated data. Even though the amount of noise is relatively small, and the fit of the model to the data is OK, the estimated input is useless. Notice the size of estimated input signal: it is a million times larger than the output. Also the elements of \hat{x} systematically alternate signs. Matlab gave a warning about the bad condition of the regression: the condition number is a little above 10^{17} !

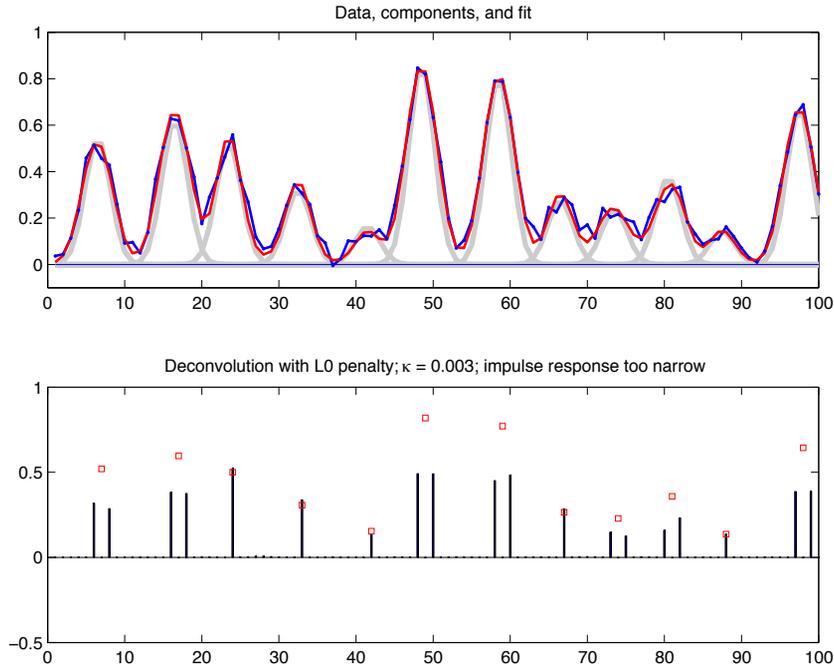


FIGURE 4. Simulated data. Upper panel: data (full blue line with dots), fit (full thick red line) and individual pulses (thick gray lines). Lower panel: deconvolution with an impulse response that is 20% too narrow.

2.2 Penalties

A simple improvement is to add the ridge penalty $\kappa \sum_j (x_j^2)$. Technically this is OK: the calculation is stable and the fit to the data generally is quite good. This is illustrated in the lower panel of Figure 1. Unfortunately, x is not very useful, because it does not look like the series of isolated sharp peaks that we expect.

An L_1 penalty, $\kappa \sum |x_j|$, gives a dramatic improvement, as Figure 2 shows. But we do not find isolated single peaks. To compute the solution in case of the L_1 penalty, I use a simple iterative procedure. It is obvious that $|x_j| = x_j^2 / |x_j|$, so if we introduce a weighted ridge penalty, with $w_j = 1/|x_j|$, we are back in the comfortable world of penalized least squares. However, we need the solution to compute the weights, so how can this be made to work? It turns out that an iterative procedure using $w_j = 1/|\tilde{x}_j|$, with \tilde{x}_j the current approximation to x_j , works well. It also turns out that using $w_j = 1/\sqrt{a + \tilde{x}_j^2}$ reduces the number of iterations. We take a small number for a , of the order of 10^6 times smaller than the expected maximum of x .

When working on this problem I made a programming error, and accidentally computed $w_j = 1/(a + \tilde{x}_j)^2$. Surprisingly this works even better. In fact this implements the L_0 penalty, the sum of non-zero elements of x , scaled by κ . Figure 2 shows that we get isolated peaks, reproducing almost perfectly the values used for the simulation.

I have not tried to optimize κ . In this application it is easy to see if a good solution is obtained, after playing a little with different values for κ . There is also a “philosophical” issue: most methods to optimize a penalty exploit prediction performance on left-out data. That is not the point here. We are interested in the input signal, not in an optimized output signal. Real instruments have a causal impulse response, meaning that it is zero for negative indices. A peak in the output always will be delayed wrt the corresponding impulse in the input. However, when visually comparing output and estimated input, it is a psychological advantage to have the peaks in both signals occur at approximately the same positions. This can be achieved by shifting the impulse response to make its peak occur near zero. In all examples this principle has been followed.

2.3 Blind deconvolution

In the simulation the impulse response is available, but for real data this is not the case. Fortunately, we do get warnings when we use a wrong impulse response, because the estimated input signal compensates the error and we will not be able to get the desired train of isolated impulses. This is illustrated in Figure 3. Generally an impulse response which is chosen too broad leads to additional negative impulses in the input, between the real one. When the impulse response is chosen too narrow, we will see close double impulses, as Figure 4 shows.

Actually, if we know x , estimating c leads to a well-conditioned regression problem. So, by alternating between estimating x and c we get an algorithm for blind deconvolution. Of course a reasonable starting estimate for c is needed, but this is not very critical and one can pick one isolated peak as a template. Figure 5 shows an example using real data from DNA sequencing. Two identifiability issues occur in blind deconvolution: 1) when we shift c p positions to the right (left) and x p positions to the left (right), we get the same y ; 2) scaling c by q and x by $1/q$ also gives the same result. The second issue is handled by dividing the estimate of c by its maximum in each iteration. In my experience, the first issue did not influence the computations. No measures were needed to prevent estimates of c from “drifting” to the left of the right.

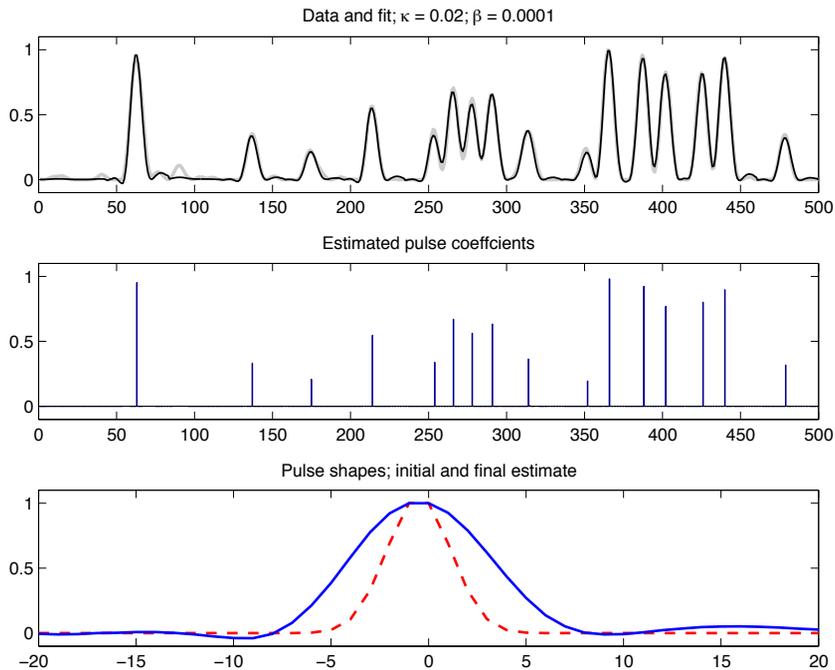


FIGURE 5. Real data from DNA sequencing. Top: data (gray) and fit (blue). Middle panel: the estimated input pulses. Bottom: the estimated impulse response (full line) and its starting estimate (broken line). A shifted impulse response has been used, to have the input peaks coincide with the top of the former.

3 Discussion

Regression with an L_0 penalty on the coefficients works well for deconvolution of pulse trains. The simple iterative algorithm finds a very good solution. This is remarkable, because the objective function is not convex, as is the case for the L_1 and L_2 penalties. Apparently there is enough structure in the problem to guide the computations in the right direction. It will be an interesting venture to develop more theoretical insight in why this is so.

On the applied side more work is needed to improve the algorithm and make it more generally useful. In real data drifting baselines can often be observed. It seems natural to model that by adding a smooth series to the model.

More challenging is the problem of gradually broadening of the impulse response. One solution is to split the data into a number of “windows”, in which the impulse response is assumed not to change much. Another approach is to introduce a power transformation of time and adjust the

exponent. In the most general case one has to assume a two-dimensional impulse response and write $y_i = \sum_j c_{ij} x_{i-j}$. Most probably additional (roughness) penalties will be needed to make this work.

In DNA sequencing four parallel series of convoluted spikes are observed, but (theoretically) a spike can occur in only one series within a certain time window. But there will be crosstalk between the channels, resulting in one large peak in the correct channel and three smaller ones in the others, for each input impulse. Li and Speed (1999) proposed a model in which a four-by-four crosstalk matrix is estimated together with the distributions of the input signals. It seems natural and feasible to extend L_0 -penalized regression to this setting, modeling crosstalk and input signals together.

References

- Brown, F.T. (2007). *Engineering System Dynamics, 2nd ed.*. CRC Press.
- Cardot, H., Koo, J-Y., Park, H.J. and A. Trubuil (2004). Boosting Diracs for Electrophoresis. *Journal of Computational and Graphical Statistics* **13**, 659–673.
- Li, L. and Speed, T.P. (1999). An estimate of the crosstalk matrix in four-dye fluorescence-based DNA sequencing. *Electrophoresis* **20**, 1433–1442.
- Li, L.M. and Speed, T.P. (2004) Deconvolution of sparse positive spikes. *Journal of Computational and Graphical Statistics* **13**, 853–870.
- Vardi, Y, and Lee, D. (1993) From image deblurring to optimal investments — maximum likelihood solutions for positive linear inverse problems. *Journal of the Royal Statistical Society B* **55**, 569–612.

A generalized random intercept log-gamma-Poisson model

Lizandra C. Fabio¹, Gilberto A. Paula¹ and Mário de Castro²

¹ Dept of Statistics, Universidade de São Paulo, Brazil, e-mail:lcfabio@ime.usp.br,

Dept of Statistics, Universidade de São Paulo, Brazil, e-mail:giapaula@ime.usp.br

² Institute of Mathematical Sciences and Computation, Universidade de São Paulo, Brazil, e-mail:mcastro@icmc.usp.br

Abstract: We propose in this work a generalized random intercept Poisson model in which the random effect distribution is assumed to follow a generalized log-gamma distribution. We express the marginal distribution for some particular parameter settings in closed-form, but in general numerical integration methods are required for deriving the marginal distribution. An application with real data is given for illustration.

Keywords: Count data; Overdispersion; Random effect models.

1 Introduction

The flexibilization of the random effect distribution in generalized linear mixed models has been investigated by some authors. For instance, Lee and Nelder (1996) have suggested this flexibilization under a hierarchical framework and more recently Molenberghs et al. (2007) have suggested a combination between gamma and normal random effects in Poisson mixed models deriving the marginal distribution.

The aim of this paper is to present an alternative distribution for the random effect in random intercept Poisson models, that is characterized by assuming a generalized log-gamma distribution for the random effect component. This distribution introduced by Prentice (1974) has as particular cases the normal and extreme value distributions and it assumes skew forms to right and left. The generalized log-gamma distribution has been widely applied in the areas Lawless, 2002 and Ortega et al., 2009). We show for some particular parameter settings that the marginal distribution assumes a closed-form expression, such as the negative binomial distribution. However, in general, numerical integration methods are required to obtain the maximum likelihood estimates. An application is given for illustration. of survival analysis and reliability (see, for instance,

2 Generalized log-gamma distribution

Let u be a random variable following a generalized log-gamma distribution. The probability density function of u (see, for instance, Lawless, 2002) is given by

$$f(u; \mu, \sigma, \lambda) = \begin{cases} \frac{c(\lambda)}{\sigma} \exp \left[\frac{(u-\mu)}{\lambda\sigma} - \frac{1}{\lambda^2} \exp \left\{ \frac{\lambda(u-\mu)}{\sigma} \right\} \right], & \text{if } \lambda \neq 0 \\ \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{(u-\mu)^2}{2\sigma^2} \right\}, & \text{if } \lambda = 0, \end{cases} \quad (1)$$

where $u, \mu, \lambda \in \mathbb{R}$, $\sigma > 0$ and are, respectively, the position, shape and scale parameters and $c(\lambda) = \frac{|\lambda|}{\Gamma(\lambda^{-2})} (\lambda^{-2})^{\lambda^{-2}}$. We will denote $u \sim \text{GLG}(\mu, \sigma, \lambda)$. The extreme value distribution is a particular case of (1), when $\lambda = 1$. For $\lambda < 0$ the pdf of u is skew to right and for $\lambda > 0$ it is skew to left. One has the moments $E(u) = \mu + \sigma \frac{\psi(\lambda^{-2}) + \log(\lambda^{-2})}{|\lambda|}$ and $\text{Var}(u) = \frac{\sigma^2 \psi'(\lambda^{-2})}{\lambda^2}$, where $\psi(\cdot)$ and $\psi'(\cdot)$ denote, respectively, the digamma and trigamma functions.

3 A generalized random intercept Poisson model

Let y_{ij} denote the j th outcome measured for the i th cluster (subject), $i = 1, \dots, n$ and $j = 1, \dots, m_i$. We will assume the following random intercept Poisson model:

- (i) $y_{ij} | u_i \stackrel{\text{indep.}}{\sim} P(\mu_{ij})$,
- (ii) $\mu_{ij} = \exp(\eta_{ij} + u_i)$ and
- (iii) $u_i \stackrel{\text{indep.}}{\sim} \text{GLG}(0, \sigma, \lambda)$,

where $\eta_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}$, $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijp})^T$ contains values of explanatory variables and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$. When $\lambda = 0$ one has the random intercept Poisson model (see, for instance, McCulloch and Searle, 2001). Let $f_{ij}(y_{ij} | u_i, \boldsymbol{\beta})$ and $f(u_i | \sigma, \lambda)$ be the pdf of $y_{ij} | u_i$ and the pdf of u_i , respectively. Then, the marginal pdf of $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_n^T)^T$, where $\mathbf{y}_i = (y_{i1}, \dots, y_{im_i})^T$, is given by

$$f(\mathbf{y} | \boldsymbol{\beta}, \sigma, \lambda) = \prod_{i=1}^n f_i(\mathbf{y}_i | \boldsymbol{\beta}, \sigma, \lambda) \quad (2)$$

with

$$f_i(\mathbf{y}_i | \boldsymbol{\beta}, \sigma, \lambda) = \int_{-\infty}^{\infty} \prod_{j=1}^{m_i} f_{ij}(y_{ij} | u_i, \boldsymbol{\beta}) f(u_i | \sigma, \lambda) du_i, \quad (3)$$

which in general does not have closed-form. However, when $\sigma = \lambda$ ($\lambda > 0$) it may be showed that (3) reduces to

$$f_i(\mathbf{y}_i | \boldsymbol{\beta}, \lambda) = \frac{\Gamma(\phi + \mathbf{y}_{i+}) \phi^\phi}{(\prod_{j=1}^{m_i} y_{ij}!) \Gamma(\phi)} \frac{\exp(\sum_{j=1}^{m_i} y_{ij} \eta_{ij})}{(\phi + \sum_{j=1}^{m_i} e^{\eta_{ij}})^{(\phi + \mathbf{y}_{i+})}}, \quad (4)$$

where $\phi = 1/\lambda^2$ and $y_{i+} = \sum_{j=1}^{m_i} y_{ij}$. In particular, when $m_i = 1$, $\forall i$, (4) reduces to the pdf of a negative binomial distribution of mean $\mu_{ij}^* = \exp(\eta_{ij})$ and dispersion parameter $\phi = \lambda^{-2}$. Thus, (4) may be interpreted as the negative multivariate binomial distribution (see, for instance, Johnson et al.'s book 1997). In particular, one has that $E(y_{ij}) = e^{\eta_{ij}}$ and $\text{Var}(y_{ij}) = c(\phi, \eta_{ij})E(y_{ij})$, where $c(\phi, \eta_{ij}) = 1 + \phi^{-1}e^{\eta_{ij}}$. The intraclass correlation between y_{ij} and $y_{ij'}$ can be expressed as, $\text{Corr}(y_{ij}, y_{ij'}) = \sqrt{e^{\eta_{ij} + \eta_{ij'}}} / \sqrt{\{e^{\eta_{ij}} + c(\phi, \eta_{ij})\} \{e^{\eta_{ij'}} + c(\phi, \eta_{ij'})\}}$.

4 Estimation and inference

For illustration we will assume $\sigma = \lambda$ ($\lambda > 0$). Thus, the marginal log-likelihood function for $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \phi)^T$, where $\phi = 1/\lambda^2$, may be expressed as

$$\begin{aligned} L(\boldsymbol{\theta}) &= \sum_{i=1}^n \left\{ \log \Gamma(\phi + y_{i+}) + \phi \log \phi - \sum_{j=1}^{m_i} \log y_{ij}! - \log \Gamma(\phi) \right. \\ &\quad \left. + \sum_{j=1}^{m_i} y_{ij} \mathbf{x}_{ij}^T \boldsymbol{\beta} - (\phi + y_{i+}) \log \left(\phi + \sum_{j=1}^{m_i} e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}} \right) \right\}. \end{aligned}$$

We use the software *R* and the *function optim* as the numerical method to maximize $L(\boldsymbol{\theta})$ and obtaining the maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\phi}$, for several fixed ϕ values, and their corresponding approximate standard errors.

5 Application

We present a biomedical example that has been described in Myers et al. (2002) in which 30 subjects (rats) have had a leukemic condition induced. Three chemotherapy type drugs were used. White (WBC) and red (RBC) blood cell counts were collected as covariates and the response is the number of cancer cell colonies. The data were collected on each subject at four different time periods.

Let $y_{ij\ell}$ denote the number of cancer cells for the i th subject at the j th period that received the treatment ℓ , $i = 1, \dots, 30$, $j = 1, 2, 3, 4$ and $\ell = 1, 2, 3$. We assume the model given by (i)-(iii) with

$$\eta_{ij\ell} = \alpha + \beta_{\ell} + \gamma_1 \text{RBC}_{ij} + \gamma_2 \text{WBC}_{ij},$$

and the restriction $\beta_1 = 0$.

The parameter estimates for $\hat{\phi} = 52$ are described in Table 1 and the results are coherent with the ones obtained by Myers et al. (2002) that applied GEE, but by this method $\text{AIC} = 1527.28$.

TABLE 1. Parameters estimates of the generalized random intercept Poisson model fitted to the rat's data.

Effect	Estimate	Std. error	<i>t</i> -value
Intercept	3.3296	0.1041	31.99
Drug2	0.2097	0.0859	2.44
Drug3	0.1958	0.0887	2.21
RBC	-0.0053	0.0048	-11.02
WBC	0.0100	0.0150	0.67
AIC	670.311		

Acknowledgments: The authors are grateful to CNPq and FAPESP, Brazil.

References

- Johnson, N.L., Kotz, S., and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. Wiley, New York.
- Lawless, J. F. (2002). *Statistical Models and Methods for Lifetime Data*, 2nd Edition. Wiley, New York.
- Lee, Y., and Nelder, J.A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B*, **58**, 619-678.
- Ortega, E.M.M., Cancho, V.G., and Paula, G.A. (2009). Generalized log-gamma regression models with cure fraction. *Lifetime Data Analysis*, **15**, 79-106.
- Prentice, R.L. (1974). A log gamma model and its maximum likelihood estimation. *Biometrika*, **61**, 539-544.
- McCulloch, C., and Searle, S. (2001). *Generalized, Linear and Mixed Models*. Wiley, New York.
- Molenbergs, G., Verbeke, V., and Demétrio, G.G.B. (2007). An extended random-effects approach to modeling repeated, overdispersed and count data. *Lifetime Data Analysis*, **13**, 513-531. *Biometrika*, **61**, 539-544.
- Myers, R.H., Montgomery, D.C., and Vining, G.G. (2002). *Generalized Linear Models with Applications in Engineering and the Sciences*, Wiley, New York.

Marginal and Conditional Akaike Information Criteria in Linear Mixed Models

Sonja Greven¹ and Thomas Kneib²

¹ Department of Biostatistics, Johns Hopkins University, USA; sgreven@jhsph.edu

² Department of Mathematics, Carl-von-Ossietzky-Universität Oldenburg, Germany; thomas.kneib@uni-oldenburg.de

Abstract: In linear mixed models, the Akaike information criterion (AIC) is often used to decide on the inclusion of a random effect. An important special case is the choice between linear and nonparametric regression models estimated using mixed model penalized splines. We investigate the behavior of two commonly used versions of the AIC, derived either from the implied marginal model or the conditional model formulation. We find that the marginal AIC is not asymptotically unbiased for twice the expected relative Kullback-Leibler distance, and favors smaller models without random effects. For the conditional AIC, it is computationally costly for large sample sizes to correct for estimation uncertainty. However, ignoring it, as is common practice, induces a bias that yields the following behavior: Whenever the random effects variance estimate is positive (even if small), the more complex model is preferred. We illustrate our results in a simulation study, and investigate their impact in modeling childhood malnutrition in Zambia.

Keywords: Kullback-Leibler information; model selection; penalized splines; random effect; variance component.

1 Introduction

Linear mixed models are increasingly used to model complex data structures. Using penalized splines, they can combine model components such as non-linear or spatial effects, interaction surfaces or varying coefficients with cluster-specific random effects. The growing flexibility of such regression models then makes the question of model selection increasingly important. The Akaike information criterion (Akaike, 1973) is often used to decide on the inclusion of random effects in linear mixed models. A common special case when using penalized splines is the decision between a linear and a nonparametric function for a covariate effect. An AIC based on the implied marginal likelihood is typically used (mAIC). Vaida & Blanchard (2005) proposed an AIC derived from the conditional model formulation (cAIC). They argue that the cAIC is more appropriate when focus is on the random effects, such as in the case of penalized splines, as the random effects are

then additional parameters that are estimated subject to a distributional constraint rather than a tool for modeling the correlation structure. However, both AIC versions are commonly used. We investigate the behavior of both for the selection of random effects in the linear mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}, \tag{1}$$

where \mathbf{X} and \mathbf{Z} are known design matrices, $\boldsymbol{\beta}$ is a fixed parameter vector, \mathbf{b} and $\boldsymbol{\varepsilon}$ are assumed to be independent, $\mathbf{b} \sim N(\mathbf{0}, \mathbf{D})$ and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

2 The marginal AIC

The AIC can be generally defined as

$$\begin{aligned} AIC = & -2\log f(\mathbf{y}|\hat{\boldsymbol{\psi}}(\mathbf{y})) + 2\mathbf{E}_{\mathbf{y}}[\log f(\mathbf{y}|\hat{\boldsymbol{\psi}}(\mathbf{y})) - \log f(\mathbf{y}|\boldsymbol{\psi}_{\mathbf{K}})] \\ & + 2\mathbf{E}_{\mathbf{y}}[\mathbf{E}_{\mathbf{z}}[\log f(\mathbf{z}|\boldsymbol{\psi}_{\mathbf{K}}) - \log f(\mathbf{z}|\hat{\boldsymbol{\psi}}(\mathbf{y}))]], \end{aligned} \tag{2}$$

where $f(\mathbf{y}|\hat{\boldsymbol{\psi}}(\mathbf{y}))$ is the maximized likelihood, and $\boldsymbol{\psi}$ are k unknown parameters with values $\boldsymbol{\psi}_{\mathbf{K}}$ minimizing the Kullback-Leibler distance (Kullback & Leibler, 1951) between the true underlying joint density $g(\cdot)$ and the family of approximating candidate models $f(\cdot|\boldsymbol{\psi})$, $\boldsymbol{\psi} \in \boldsymbol{\Psi}$,

$$K(f_{\boldsymbol{\psi}}, g) = \int \{\log(g(z)) - \log(f_{\boldsymbol{\psi}}(z))\}g(z)dz = \mathbf{E}_z[\log(g(z)) - \log(f_{\boldsymbol{\psi}}(z))].$$

$K(f_{\boldsymbol{\psi}}, g)$ can be viewed as a measure of distance between $g(\cdot)$ and $f(\cdot|\boldsymbol{\psi})$, $\boldsymbol{\psi} \in \boldsymbol{\Psi}$. As the AIC is unbiased for twice the expected relative Kullback-Leibler distance, minimizing (2) can be seen as minimizing the average distance of an approximating model to the underlying truth.

In standard cases, certain regularity conditions are fulfilled, including that observations are independent and identically distributed, and the parameter space (up to a change of coordinates) is R^k . Then, the last two terms in (2) reduce to $2k$ asymptotically. This is the AIC commonly used.

The marginal AIC (mAIC) in the linear mixed model uses the likelihood of the implied marginal model $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ with $\mathbf{V} = \mathbf{I}_n + \mathbf{Z}\mathbf{D}\mathbf{Z}'$. The number of estimable parameters then is $p + q$, with $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ and q the number of unknown parameters $\boldsymbol{\theta}$ in \mathbf{V} . Thus, the mAIC is defined as

$$mAIC = -2\log(f(\mathbf{y}|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})) + 2(p + q).$$

Now, we can show (Greven & Kneib, 2008) that due to the marginal correlation structure in \mathbf{y} in (1) and the constraints on $\boldsymbol{\theta}$ (variances have to be non-negative, and more generally, \mathbf{D} has to be positive semi-definite), the last two terms in (2) are smaller than $2(p + q)$ as well as not independent of the true values in $\boldsymbol{\theta}$. Consequently, the mAIC is positively biased, and favors smaller models without random effects.

3 The conditional AIC

Vaida & Blanchard (2005) define the conditional AIC (cAIC) as

$$cAIC = -2\log(f(\mathbf{y}|\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}, \hat{\boldsymbol{\theta}})) + 2(\rho + 1),$$

where $f(\mathbf{y}|\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}, \hat{\boldsymbol{\theta}})$ is the maximized conditional likelihood (conditioning on \mathbf{b} as well as on $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$), $\hat{\mathbf{b}}$ is the best linear unbiased predictor of \mathbf{b} , $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ are the maximum likelihood (ML) or restricted maximum likelihood (REML) estimates of $(\boldsymbol{\beta}, \boldsymbol{\theta})$, and

$$\rho = \text{trace} \left(\left(\begin{array}{cc} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{D}_*^{-1} \end{array} \right)^{-1} \left(\begin{array}{cc} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} \end{array} \right) \right).$$

This definition of ρ corresponds to the trace of the hat matrix, and is connected to the effective degrees of freedom definition known from smoothing. The authors assume $\mathbf{D}_* = \sigma^{-2}\mathbf{D}$ to be known, but suggest using $\hat{\rho}$ with estimated \mathbf{D}_* otherwise, arguing that the difference is negligible for large n . We will call this the conventional or simplified cAIC in the following. Liang et al. (2008) propose a corrected cAIC, accounting for estimation of \mathbf{D}_* . For known σ^2 , they replace ρ by $\Phi_0 = \text{trace}(\partial\hat{\mathbf{y}}/\partial\mathbf{y})$, where $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}\hat{\mathbf{b}}$. For unknown σ^2 , the effective degrees of freedom Φ_1 involve even second derivatives,

$$\Phi_1 = \frac{\tilde{\sigma}^2}{\hat{\sigma}^2} \text{trace} \left(\frac{\partial\hat{\mathbf{y}}}{\partial\mathbf{y}} \right) + \tilde{\sigma}^2 (\hat{\mathbf{y}} - \mathbf{y})' \frac{\partial\hat{\sigma}^{-2}}{\partial\mathbf{y}} + \frac{1}{2} \tilde{\sigma}^4 \text{trace} \left(\frac{\partial^2\hat{\sigma}^{-2}}{\partial\mathbf{y}\partial\mathbf{y}'} \right),$$

where $\tilde{\sigma}^2$ is an estimate for the true error variance. As these derivatives are not available in closed form, numerical approximations using n respectively $2n$ additional model fits have to be used. This can be prohibitive in large samples. In our application ($n = 1600$, 64 models to compare), we estimated the necessary computation time to be about 110 days. As the authors in their simulations find only small differences between conventional and corrected cAIC, we investigate whether the often used simplified cAIC is a computationally feasible alternative - especially when n is large. In this case, the computational cost of the corrected cAIC can be too high, and consistent estimators should yield precise variance estimates. Unlike Liang et al. (2008a), who concentrated on estimating the effective degrees of freedom, we focus on the performance for differentiating between zero and non-zero random effects variances.

Surprisingly, we can show (Grevén & Kneib, 2008) that ignoring estimation uncertainty in \mathbf{D}_* for the simplified cAIC results in the following interesting behavior (for simplicity, we focus on the case of one unknown variance component, i.e. $\mathbf{D} = \tau^2\boldsymbol{\Sigma}$ with known $\boldsymbol{\Sigma}$): When $\hat{\tau}^2 = 0$, the cAICs of the models including and excluding \mathbf{b} agree, i.e. there is a tie. When $\hat{\tau}^2 > 0$, the cAIC prefers the larger model including \mathbf{b} , regardless of the size of

$\hat{\tau}^2$. The simplified cAIC thus is not a useful decision rule, as it does not give guidance on when an estimated variance is large enough to warrant inclusion of the random effect in the model, or small enough to justify exclusion of the random effect from the model.

The principal difficulty of the simplified cAIC is that the degrees of freedom in the cAIC are estimated from the same data as the model parameters. This leads to a bias that results in a preference for larger models. This behavior has its analogy in the AIC itself. Use of the maximized log-likelihood for model choice would always result in the largest model being chosen. The underlying over-optimism in the model fit is due to the parameter estimates being obtained from the same data which is the argument of the log-likelihood. The AIC corrects for this bias and is a truly predictive quantity. A similar mechanism is at play here. While the correct bias correction term in our case cannot be derived analytically, Liang et al. (2008a) circumvent the problem using numerical derivatives. In a sense, their bias correction term is measuring the sensitivity of results to new data, similar in spirit to other predictive criteria such as generalized cross validation (GCV). Unfortunately, this comes at the price of computational complexity (and some numerical instability) that is comparable to leave-one-out cross validation. More work is clearly needed here.

4 Simulations

First, we compare a linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with the nonparametric regression model $y_i = m(x_i) + \varepsilon_i$, modeled using penalized splines in the mixed model framework. Thus, the comparison corresponds to selecting a random effect modeling deviations of $m(\cdot)$ from linearity. The true functions are chosen as (see Figure 1)

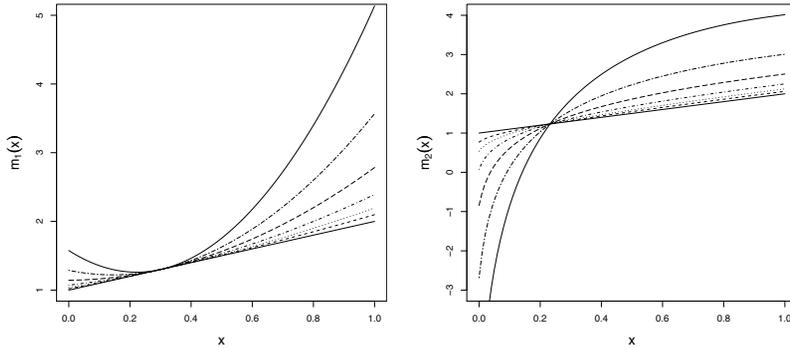
$$\begin{aligned} m_1(x) &= 1 + x + 2d(0.3 - x)^2, \\ m_2(x) &= 1 + x + d(\log(0.1 + 5x) - x), \\ m_3(x) &= 1 + x + 0.3d(\cos(0.5\pi + 2\pi x) - 2x) \end{aligned}$$

with varying non-linearity parameter $d = 0, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2$, where $d = 0$ corresponds to linearity. The sample size is taken as $n = 30, 50, 100, 200$. The error variance is set to $\sigma^2 = 1$, and x is chosen equidistantly from the interval $[0, 1]$.

Second, we compare a random intercept and a common intercept model, varying random intercept variance, number and size of clusters. In case of a tie between models, we intrinsically decide on the smaller model.

The simplified cAIC gives a much larger proportion of decisions for the larger model than the mAIC, with the corrected cAIC in-between (Figure 2). While the AIC for nested models in standard settings corresponds to a likelihood ratio test with asymptotic level $\alpha = 0.157$, α is much smaller for the mAIC (as low as 0.01 in our simulations), much larger for the simplified cAIC (up to 0.49), and more similar for the corrected cAIC (0.07 to 0.40).

FIGURE 1. Functions $m_1(\cdot)$ and $m_2(\cdot)$ for different values of the non-linearity parameter d .



As predicted from our theoretical results, the simplified cAIC chooses the larger model when $\hat{\tau}^2 > 0$, and gives a tie when $\hat{\tau}^2 = 0$. Thus, α here simply corresponds to the proportion of non-zero variance estimates given a true zero variance, which for penalized splines is about 20% for ML estimation, and more than 35% for REML estimation, and which approaches 50% for the random intercept model.

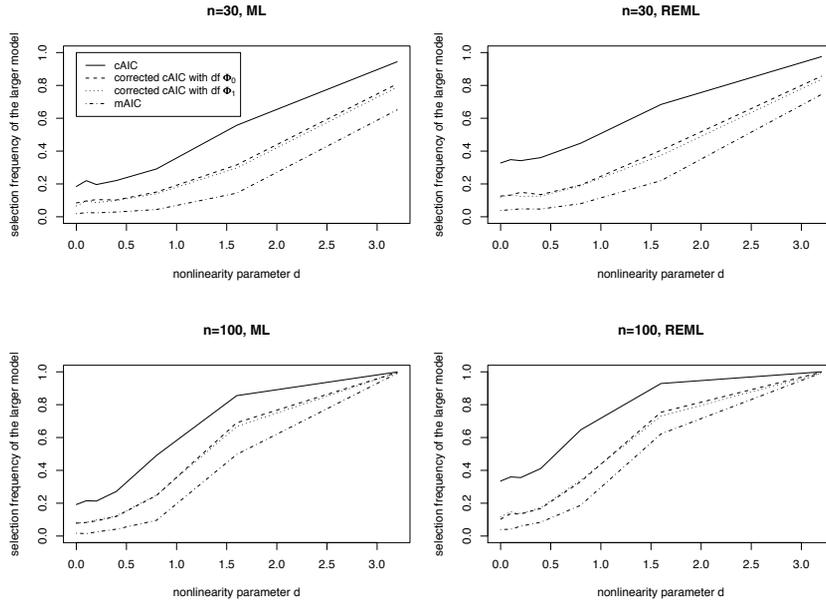
In contrast, the mAIC does not show this behavior, and in particular never yields equality under the linear and the non-linear model due to the additional parameter count for the variance parameter. The corrected cAIC with $\Phi_0 + 1$ still results in a large number of ties, which disappear when using Φ_1 .

The corrected cAIC often favors the more complex model even when $\hat{\tau}^2 = 0$ due to numerical problems. Especially for Φ_1 using second derivatives, the numerical approximation fails in some cases, resulting in spurious estimated degrees of freedom. Overall, $\Phi_0 + 1$ approximates Φ_1 rather well (see Figure 2), but is numerically much more stable.

5 Childhood malnutrition in Zambia

We investigate implications of our theoretical findings for model choice in practice. We are interested in modelling the Z-score, measuring chronic undernutrition (stunting) as insufficient height for age, for 1600 children from the 1992 Zambia Demographic and Health Survey. The available predictors were 1) categorical/binary: child's gender, mother's employment status and education 2) spatial: residential district and 3) continuous: duration of breastfeeding, child's age, mother's age, height and body mass index. Due to computational cost of the corrected cAIC in our large data set, we focused only on the mAIC and the simplified cAIC for selecting a random

FIGURE 2. Selection frequencies of the larger, non-linear model in our simulations for function $m_1(\cdot)$.



district intercept, and linear or non-linear effects for the continuous variables. Categorical and binary variables were modeled parametrically, giving a total of 64 models to choose from.

To illustrate our findings in a simple example, consider the model with height of the mother as the single predictor. Using maximum likelihood estimation, the estimated effect is linear (Figure 3). This results in a tie for the cAIC, while the mAIC clearly prefers the smaller linear model due to the additional parameter count for the variance parameter. Using REML estimation, the estimated effect is slightly non-linear. While the mAIC still prefers the smaller, linear model, the cAIC as expected chooses the larger, non-linear model, despite the estimated non-linearity being quite small.

For the overall comparison of all 64 models, let a tie in the cAIC be indicative of a choice of the simpler model. Then, cAIC and mAIC for both ML and REML agree on the overall best model including a random district intercept, linear effects for age, height and body mass index of the mother, and non-linear effects of child’s age and duration of breastfeeding (Figure 4). As mAIC and simplified cAIC are biased in opposite directions, agreement between the two indicates optimality of this final model.

FIGURE 3. Estimated effect of height of the mother (in cm) on the Z-score measuring chronic undernutrition of children in Zambia.

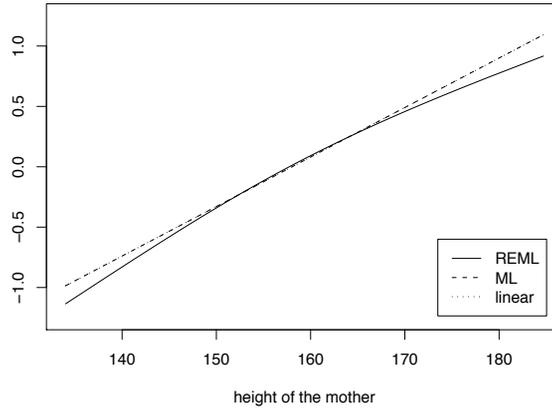
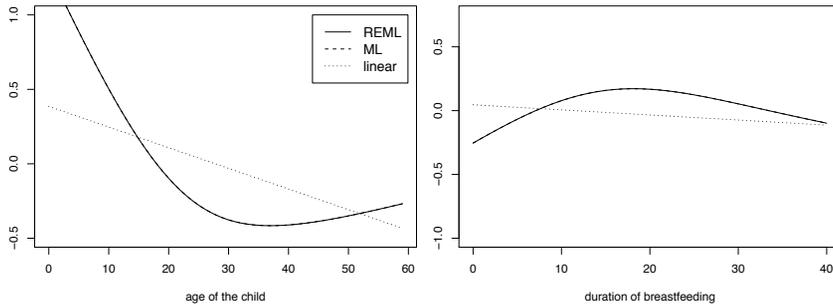
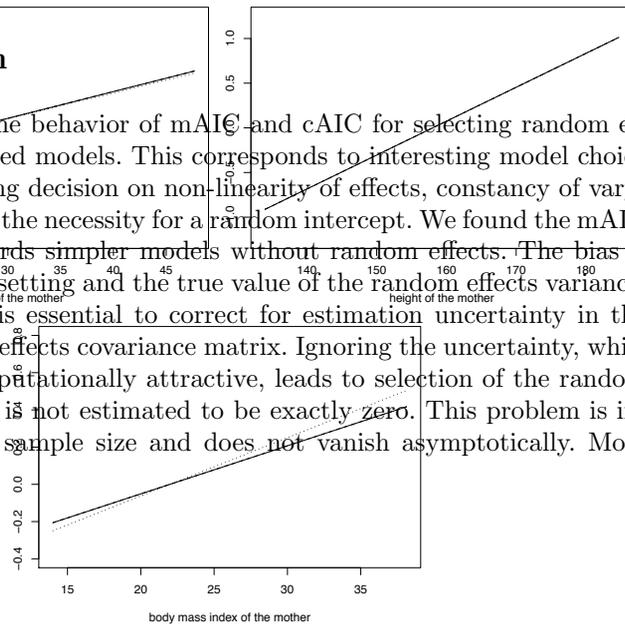


FIGURE 4. Estimated effects of age of the child and duration of breastfeeding (in months) on the Z-score measuring chronic undernutrition of children in Zambia.



6 Discussion

We investigated the behavior of mAIC and cAIC for selecting random effects in linear mixed models. This corresponds to interesting model choice questions, including decision on non-linearity of effects, constancy of varying coefficients, or the necessity for a random intercept. We found the mAIC to be biased towards simpler models without random effects. The bias is dependent on the setting and the true value of the random effects variance. For the cAIC, it is essential to correct for estimation uncertainty in the unknown random effects covariance matrix. Ignoring the uncertainty, while common and computationally attractive, leads to selection of the random effect whenever it is not estimated to be exactly zero. This problem is independent of the sample size and does not vanish asymptotically. More



research is needed to obtain numerically feasible and robust versions of the corrected cAIC, and to extend methodology to generalized linear mixed models.

For a longer working paper on our results including proofs, please see Greven & Kneib (2008).

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: *2nd International Symposium on Information Theory*. 267-281, Akademiai Kiado.
- Greven, S. and Kneib, T. (2008). On the Behavior of Marginal and Conditional Akaike Information Criteria in Linear Mixed Models. *Johns Hopkins University, Department of Biostatistics Working Papers*, Paper 179. <http://www.bepress.com/jhubiostat/paper179/>
- Kullback, S. and Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, **22**, 79–86.
- Liang, H., Wu, H. and Zou, G. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika*, **95**, 773–778.
- Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, **92**, 351-370.

Exponentially Weighted Poisson Models

Linda M. Haines¹ and Kerry L. Leask¹

¹ Department of Statistical Sciences, University of Cape Town, Private Bag X3, Rondebosch 7701, South Africa. *e-mail* : linda.haines@uct.ac.za

Abstract: This paper is concerned with examining formally the properties of the exponentially weighted Poisson distribution and with the usefulness of this distribution in modelling simple count data and data obtained from a Wadley's problem setting.

Keywords: Count data; Overdispersion; Underdispersion; Wadley's problem.

1 Introduction

Count data are often overdispersed or, more occasionally, underdispersed relative to the benchmark Poisson distribution and a wide range of Poisson-based and other distributions have been developed in order to address this issue. In particular Ridout and Besbeas (2004) introduced the exponentially weighted Poisson (EWP) distribution which weights Poisson probabilities with the exponential of minus the scaled distance of the count from the Poisson parameter and demonstrated the scope and flexibility of their model by means of a series of examples. The aim of the present study is to examine the properties of the EWP distribution more formally, to highlight the advantages and disadvantages that accrue and to probe more broadly the practical usefulness of the distribution in modelling over- and underdispersed count data.

2 Distribution

2.1 Properties

Suppose that the random variable Y follows the two-parameter *EWP* distribution of Ridout and Besbeas (2004). Then the probability mass function (p.m.f.) of Y is given by

$$Pr(Y = y) = \frac{e^{-\tau} \tau^y e^{-\theta|y-\tau|}}{y! W}, y = 0, 1, 2, \dots$$

where τ is the Poisson parameter and θ is a parameter incorporating over- or underdispersion. The term W represents the normalizing constant, that

is $W = \sum_{y=0}^{\infty} \frac{e^{-\tau} \tau^y e^{-\theta|y-\tau|}}{y!}$. Now, by some straightforward algebra, W can be re-expressed as the finite sum

$$W = \sum_{y=0}^{\lfloor \tau \rfloor} \frac{e^{-\tau} \tau^y e^{\theta(y-\tau)}}{y!} + e^{\tau(\theta-1+e^{-\theta})} - \sum_{y=0}^{\lfloor \tau \rfloor} \frac{e^{-\tau} \tau^y e^{-\theta(y-\tau)}}{y!}$$

where $\lfloor \tau \rfloor$ denotes the largest integer less than or equal to τ and any sum for which the upper limit is strictly less than the lower limit is taken as 0. Thus the probabilities $Pr(Y = y), y = 0, 1, \dots$, can be calculated explicitly. This result is in contrast to other weighted Poisson distributions such as the Conway-Maxwell-Poisson (CMP) for which calculations of probabilities necessitate the approximating of infinite sums (Shmueli, Minka, Kadane, and Boatwright, 2005). More crucially, it greatly facilitates the calculation of moments and likelihoods. For example the mean of Y is readily derived as the finite sum

$$E(Y) = \frac{\tau e^{-\tau(1+\theta)+\theta}}{W} \sum_{y=0}^{\lfloor \tau \rfloor - 1} \frac{(\tau e^{\theta})^y}{y!} + \frac{\tau e^{\tau(\theta-1)-\theta}}{W} \left[e^{\tau e^{-\theta}} - \sum_{y=0}^{\lfloor \tau \rfloor - 1} \frac{(\tau e^{-\theta})^y}{y!} \right].$$

Clearly the *EWP* distribution reduces to the Poisson when $\theta = 0$. For $\theta > 0$, the weight $exp(-\theta|y - \tau|)$ in the p.m.f. for Y induces thin tails, thereby allowing the *EWP* distribution to accommodate underdispersion. Conversely, for $\theta < 0$ the *EWP* distribution has fatter tails and indeed in the extreme, as θ gets larger, approaches bimodality.

2.2 Estimation and inference

Suppose now that a random sample of n observations y_1, \dots, y_n is taken from the two-parameter *EWP* distribution. Then the likelihood function is the joint p.m.f. of these observations and the log-likelihood is thus given by

$$\ell(\tau, \theta) = \sum_{i=1}^n y_i \ln \tau - n\tau - \sum_{i=1}^n \ln y_i! - \sum_{i=1}^n \theta |y_i - \tau| - n \ln W.$$

The presence of absolute values $|y_i - \tau|$ in this expression leads to discontinuities in the log-likelihood function at integer values of τ . Thus the function $\ell(\tau, \theta)$ is not differentiable with respect to τ and maximization of the log-likelihood with respect to the parameters τ and θ is not straightforward. One approach to this problem, and the one used here, is to fix the parameter τ at an integer value, to maximize the resultant log-likelihood with respect to θ and then to repeat the process for a suitable range of integer values of τ . Inference for the unknown parameters is similarly awkward and bootstrapping procedures are investigated in an attempt to address the problems.

2.3 Examples

Two examples, one the overdispersed quarterly sales data presented in Shmeuli et al (2005) and the other data simulated from a mixture of two Poissons, are introduced in order to appraise the practical usefulness of the results outlined above. In each case the goodness-of-fit of the *EWP* distribution is assessed by examining the residuals and by invoking Pearson's χ^2 statistic and comparisons with other models are also made. It is worth noting that the *EWP* distribution provides a remarkably good fit to the simulated two-Poisson mixture data.

3 Wadley's problem

3.1 Setting and properties

Consider now Wadley's problem within the context of a dose-response study. Specifically, suppose that Y , the number of organisms that survive exposure to varying doses of a drug, is observed and that the number initially treated, N , is unknown. Then Y given $N = n$ can be modelled as a binomial with probability p of survival and the unknown N can be taken to follow an appropriate count distribution. The probability p is usually modelled with the logit link function $\text{logit}(p) = \beta_0 + \beta_1 x$, where β_0 and β_1 are unknown regression parameters and x represents dose or log dose. Thus, if the *EWP* distribution is used to model N , then the marginal p.m.f. of Y follows immediately, with W representing the appropriate normalizing constant, as

$$\begin{aligned} Pr(Y = y) = & \frac{2(\tau p)^y e^{-\tau}}{y! W} \sum_{k=0}^{\lfloor \tau \rfloor - y} \frac{[\tau(1-p)]^k}{k!} \sinh[\theta(k+y-\tau)] \\ & + \frac{(\tau p e^{-\theta})^y e^{\tau(\theta-1+(1-p)e^{-\theta})}}{y! W} \end{aligned}$$

for $y < \tau$ and as

$$Pr(Y = y) = \frac{(\tau p e^{-\theta})^y e^{\tau(\theta-1+(1-p)e^{-\theta})}}{y! W}$$

for $y > \tau$. The variable Y is then said to follow a binomial-EWP distribution with parameters τ, β_0, β_1 and θ . Crucially the p.m.f. of the binomial-EWP distribution includes only finite sums and can therefore be calculated exactly. Explicit expressions for the moments of this distribution can also be found but are rather cumbersome.

3.2 Estimation and inference

Suppose now that data from a Wadley's problem setting are collected for a set of controls and for a range of doses. Then the log-likelihood can be

written down in a straightforward manner based on the expressions for the probabilities $Pr(Y = y)$ presented above. However the same problems with respect to maximizing the log-likelihood as those encountered for the basic *EWP* distribution occur. Thus integer values of τ are again invoked in order to obtain suitable maximum likelihood estimates of the parameters and approaches to inference based on Wald intervals, profile likelihoods and bootstrap methods are investigated.

3.3 An example

Baker, Pierce and Pierce (1980) present data on the growth of unicellular bacteria under exposure to differing concentrations of chemicals, with only the numbers of surviving bacteria recorded. This data set is now used to illustrate the properties and features of the binomial-*EWP* model outlined above. Specifically, parameters are estimated and inferences drawn. In addition the results are compared with those obtained by fitting models for which N follows a Poisson and a negative binomial distribution and models for which N is Poisson and overdispersion is introduced through the probability of survival p .

4 Discussion

There are a number of immediate extensions to the work reported here and some of these will be discussed. In particular, Ridout and Besbeas (2004) also proposed a three-parameter *EWP* distribution and the flexibility in modelling so introduced is examined. In addition surprising and somewhat serious problems in the fitting of *EWP* models to regression count data have been encountered and these issues are explained and explored.

Acknowledgments: The authors would like to thank the University of Cape Town, the University of KwaZulu-Natal, and the National Research Foundation, South Africa, for financial support.

References

- Baker R.J., Pierce C.B., and Pierce J.M. (1980). Wadley's problem with controls. *GLIM Newsletter*, **3**, 32-35.
- Ridout M.S., and Besbeas P. (2004). An empirical model for underdispersed count data. *Statistical Modelling*, **4**, 77-89.
- Shmueli G., Minka T.P., Kadane J.B., and Boatwright P. (2005). A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution. *Applied Statistics*, **54**, 127-142.

A Bayesian Approach for Evaluating Drug Efficacy using Fecal Egg Count Data

Bret M. Hanlon¹, Anand N. Vidyashankar¹, Stig L. Petersen², Ray M. Kaplan³, and Martin K. Nielsen⁴

¹ Department of Statistical Sciences, Cornell University, Ithaca, New York, USA

² EquiLab Laboratory, Slangerup, Denmark

³ Department of Infectious Diseases, University of Georgia, Georgia, USA

⁴ Department of Large Animal Sciences, University of Copenhagen, Denmark

Abstract: An important problem in equine parasitology is the treatment of horses for parasitic worms with anthelmintic drugs and the evaluation of drug efficacy. The current practical gold standard for monitoring treatment efficacy is based on the fecal egg count data. This paper presents a Bayesian procedure for analyzing such data. The procedure is applied to a 64 farm study consisting of 614 horses treated with the anthelmintic pyrantel.

Keywords: Bayesian statistics; anthelmintic resistance; equine parasitology; fecal egg count data

1 Introduction

Anthelmintic resistance in gastrointestinal nematode parasites is an emerging problem in equine parasitology (Kaplan, 2002, 2004). The first step towards addressing resistance is to estimate the efficacy of the given drug. The theoretical definition of drug efficacy involves counting the number of killed and live worms after treatment. However, these data can only be obtained by sacrificing the animal, which is not practical on a farm. The practical gold standard for assessing treatment efficacy is based on fecal egg count (FEC) data. These data are obtained by measuring the number of eggs in sample of feces, pre- and post-treatment. Frequentist statistical methods for the evaluation of efficacy using FEC data are studied in Vidyashankar et al.(2007).

To motivate our investigation, we discuss some recent developments in the treatment of equine nematodes (see Nielsen et al., 2006, 2008, for more details). FEC studies typically involve counts of cyathostomin nematodes (CNs), which are presently considered as the most dangerous parasitic threat to pasture horses. There is new evidence which suggests the need to monitor other equine nematodes, in particular *Strongylus vulgaris* (SV). SV is a highly pathogenic nematode and is capable of causing severe disease in infected horses; however, years of intensive anthelmintic treatment

has greatly decreased its prevalence. Recent studies have shown an increase of SV on Danish horse farms, which is likely related to new national policies making anthelmintics available by prescription only. The legislation, introduced in 1999, is aimed at delaying the development of anthelmintic resistance in CNs.

The present paper is concerned with analysis of data that has information on two types of nematodes, namely SV and CN. We use the term *drug efficacy* to mean the effectiveness of the treatment for eliminating CNs. Our approach is based on a Bayesian hierarchical model to address the following two scientific questions: (i) does the presence of SV influence treatment efficacy and (ii) is there evidence of reduction in anthelmintic efficacy. We apply our procedure to analyze FEC data from 64 Danish horse farms treating with the anthelmintic pyrantel. Our results suggest that SV presence is not associated with pyrantel efficacy. Although overall efficacy of pyrantel was high, some of the horse farms exhibit low efficacies.

The next section discusses the data from the Danish horse farm study. Section 3 presents our Bayesian model for analyzing this data; the data analysis is described in Section 4. Finally, some concluding remarks are given in Section 5.

2 Data

Individual fecal samples were taken from 1644 horses on 64 farms in Denmark. In this study, horses with pre-treatment fecal egg counts of 200 or higher were treated with pyrantel. Each farm had at least six horses receiving the treatment, with a total of 614 horses receiving treatment. Our data analyses are based on these 614 horses. Larval cultures were also performed on all pre-treatment samples for identification of SV. Of the 64 farms, 30 farms had at least one horse infected with SV, we refer to these farms as *SV positive*.

3 Model

3.1 General Setting

Following Vidyashankar et al. (2007), we use a generalized linear mixed model to describe the data and efficacy. More precisely, let $(N_{i,j,k}, X_{i,j,k})$ denote the pre-treatment egg count and post-treatment reduction (*precount*–*postcount*) for the j^{th} horse on farm i with SV status k ; $k = 0$ represents SV negative farms, while $k = 1$ represents SV positive farms. We assume that

$$\begin{aligned} X_{i,j,k} | N_{i,j,k}, p_{i,k} &\sim \text{Bin}(N_{i,j,k}, p_{i,k}) \\ p_{i,k} &\sim G(\theta_k), \end{aligned}$$

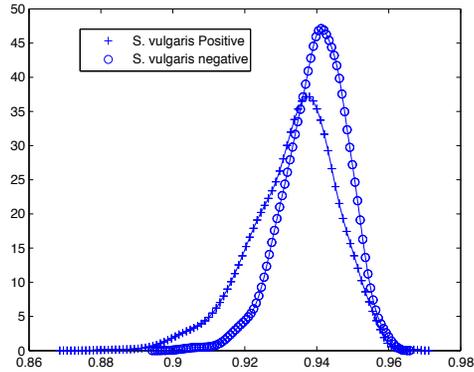


FIGURE 1. KDEs for the posterior distributions of η_0 and η_1 . The KDE for the SV positive farms is labeled with +, while the KDE for the SV negative farms is labeled with o.

where G represents the random effect distribution. Since our methodology is based on a Bayesian approach, to complete the description of the model we specify a prior for θ_k .

3.2 Implementation

For our data analysis we set $G(\theta_k)$ to be the beta distribution with parameters α_k, β_k . That is, $p_{i,k} | (\alpha_k, \beta_k) \sim \text{Beta}(\alpha_k, \beta_k)$. Furthermore we model α_k and β_k as independent $\text{gamma}(1, .001)$ random variables.

We use WinBUGS (Spiegelhalter et al., 2003) to approximate the posterior distribution for η_k , where $\eta_k \equiv \frac{\alpha_k}{\alpha_k + \beta_k}$ is the conditional expected value of $G(\theta_k)$. Our samplers involve a single chain of 5000 iterations, with a 1000 sample burn-in. Of the remaining 4000 samples, we used a thinned 2000 samples for posterior estimation and inference. The standard diagnostic checks (not shown) suggest proper convergence of all the samplers considered; for diagnostic purposes, we additionally ran three chains with different starting values to check for proper mixing.

4 Data Analysis

We begin by addressing the impact of SV on pyrantel efficacy. Figure 1 displays kernel density estimates (KDEs) for the posterior distributions of η_0 and η_1 . A KDE for the posterior distribution of $\eta_1 - \eta_0$ is given in Figure 2. It appears that the presence of SV does not greatly impact drug

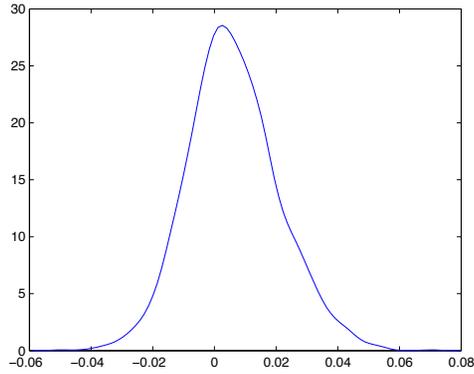


FIGURE 2. A KDE of the posterior distribution of $\eta_1 - \eta_0$.

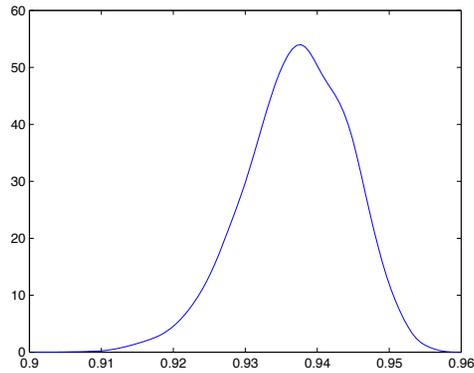


FIGURE 3. A KDE of the posterior distribution of η (this analysis is based on data from all 64 farms).

efficacy. In fact, the 95% credible region (and posterior mean) for η_0, η_1 and $\eta_1 - \eta_0$ are given by $[0.9204, 0.9550]$ (0.9398), $[0.9057, 0.9538]$ (0.9337), and $[-0.0370, 0.0206]$ (-0.0061), respectively.

We next study the question of drug efficacy. Figure 3 plots a KDE for the posterior distribution of η , combining the data for all 64 farms. The posterior mean and 95% posterior credible region for η are 0.9372 and $[0.9218, 0.9499]$. These numbers coincide with the presumed efficacy of pyrantel for non-resistant worms, and thus do not suggest evidence of re-

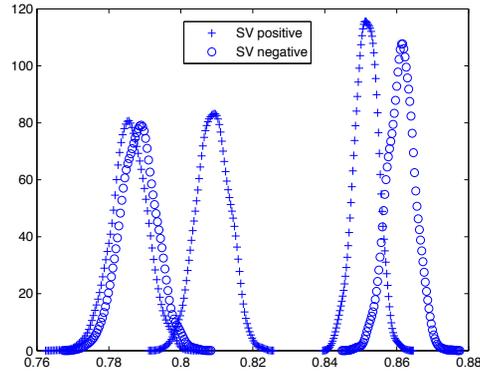


FIGURE 4. KDEs for the posterior distributions of the farm level efficacy (p_i) from the five farms with the lowest posterior means. The SV positive farms are labeled with +, the SV negative farms are labeled with o .

sistance. There are, however, five farms which have relatively low observed efficacies. KDEs for the posterior distributions of these farms' efficacies are given in Figure 4. It is possible that these low observed efficacies are caused by the presence of “outlier horses” or are due to the inherent variability present in the mechanism for obtaining FEC data (Kaplan, 2004). More detailed analysis is needed for addressing this issue.

5 Conclusions

We developed a Bayesian procedure for the analysis of FEC data and for addressing the related issue of treatment efficacy. Additionally, we provide an extension of this procedure to address the two sample problem, namely evaluating the difference in drug efficacy between two treatment groups. We applied this methodology to a study of 64 Danish horse farms, which can be categorized as SV-positive or SV-negative. Our analysis suggests that SV was not associated with pyrantel efficacy. Although overall efficacy of pyrantel was high, some of the horse farms exhibit low efficacies.

Acknowledgments: ANV's research is supported in part by a grant from NSF DMS 000-03-07057 and also by grants from the NDCHealth Corporation.

References

- Kaplan, R. (2002). Anthelmintic resistance in nematodes of horses. *Veterinary Research*, **33**.
- Kaplan, R. (2004). Drug resistance in nematodes of veterinary importance: a status report. *Trends in Parasitology*, **20**.
- Nielsen, M. K., Monrad, J., and Olsen, S.N. (2006). Prescription-only anthelmintics: a questionnaire survey of strategies for surveillance and control of equine strongyles in Denmark. *Veterinary Parasitology*, **135**, 4755.
- Nielsen, M. K., Peterson, D. S., Monrad, J., Thamsborg, S. M., Olsen, S. N. and Kaplan, R. M. (2008). Detection and semi-quantification of strongylus vulgaris DNA in equine faeces by real-time quantitative PCR. *International Journal for Parasitology*, **38**.
- Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2003). WinBUGS User Manual Version 1.4. *MRC Biostatistics Unit*.
- Vidyashankar, A. N., Kaplan, R., and Chan, S. (2007). Statistical approach to measure the efficacy of anthelmintic treatment on horse farms. *Parasitology*, **134**.

Randomly stopped sum models: a hydrological application

Gillian Z. Heller¹, D. Mikis Stasinopoulos², Robert A. Rigby²
and Floris F. van Ogtrop³

¹ Department of Statistics, Macquarie University, Sydney, Australia (gillian.heller@mq.edu.au)

² Statistics, OR, and Mathematics (STORM) Research Center, London Metropolitan University, U.K.

³ Faculty of Agriculture, Food & Natural Resources, University of Sydney, Australia

Abstract: A regression framework for the modelling of a response variable which is an accumulation of individual positive amounts is introduced. Sub-classes of this framework based on the level of detail observed are defined. One of these sub-classes is developed, and a hydrological application is presented.

Keywords: randomly stopped sums; zero-adjusted distributions; BCT distribution; streamflow.

1 Introduction

Response variables which can either be zero or a continuous positive quantity are commonly encountered. Examples are: individual alcohol consumption per week; amount claimed on insurance policies per year; rainfall per day. Using rainfall as an example, in a fixed period the rainfall amount may be zero, or there may be one or more rainfall episodes. The outcome of interest here is the total amount of rainfall for the period. With rainfall the individual episodes are typically difficult to observe and record, but in other situations the episodes are well defined. For example, a single claim on an insurance policy is an episode. There can be more than one claim, and usually the quantity of interest is the expected total amount claimed over a fixed period. In all the above situations the response variable (if it is not zero) is an accumulation of individual amounts. We propose a unified regression framework for modelling a response variable which is the total amount. In this talk we develop one of the sub-classes of this framework, and present an application to hydrological data.

2 The statistical framework

Let Y be the total amount over the period and C the number of episodes within this period. Hence Y is 0 if $C = 0$; if $C > 0$, Y is a sum of random variables Z_j , $j = 1, \dots, C$, where the number of terms C in the sum is random:

$$Y = \begin{cases} 0 & \text{if } C = 0 \\ \sum_{j=1}^C Z_j & \text{if } C = 1, 2, \dots \end{cases}$$

Being the sum of a random number of terms, Y is called a *randomly stopped sum* (Stuart and Ord 1994). Therefore we adopt the term *randomly stopped models* (RSMs) to describe models having the above type of structure as response variable. We will be confining attention to the usual case where the Z_j 's are positive and continuous; usually the distribution of Z_j is right-skewed. There are three distinct cases arising in practice:

Case 1: Y is observed, but neither C nor the individual Z_j 's are observed.

Case 2: Y and C are observed, but not the Z_j 's.

Case 3: C and the Z_j 's are observed.

In this talk we consider Case 1 in detail.

3 Case 1: Zero-adjusted models

This situation arises with rainfall and streamflow data, where it is impractical to observe or record the individual episodes. Let $f_C(c; \theta_1)$ be the probability function for C and $f_{Y|C}(y|c; \theta_2)$ the conditional distribution function for Y given C . Let $\theta = (\theta_1, \theta_2)$. The marginal distribution of Y is given by:

$$\begin{aligned} f_Y(y; \theta) &= \sum_{c=0}^{\infty} f_{Y|C}(y|c; \theta_2) f_C(c; \theta_1) \\ &= \begin{cases} f_C(0; \theta_1) & \text{if } y = 0 \\ \sum_{c=1}^{\infty} f_{Y|C}(y|c; \theta_2) f_C(c; \theta_1) & \text{if } y > 0. \end{cases} \end{aligned} \quad (1)$$

Model 1.1

Here we model the distribution of C and of Y given C as in (1). As C is not observed, the marginal distribution of Y in (1) is fitted directly. For example if C is assumed to have a Poisson distribution and the Z_j 's independent gamma distributions, then the marginal distribution of Y is the Tweedie distribution (Jørgensen and de Souza 1994, Smyth and Jørgensen 2002).

Model 1.2

Here we model the non-zero distribution of Y directly, i.e. the distribution of $T = \sum_{j=1}^C Z_j$ for $C > 0$. This leads to a *zero-adjusted* distribution for Y , that is, a finite mixture distribution having the following form:

$$f_Y(y; \theta, \pi) = \begin{cases} 1 - \pi & \text{if } y = 0 \\ \pi f_T(y; \theta_2) & \text{if } y > 0 \end{cases} \quad (2)$$

where $0 < \pi < 1$, and θ_2 is the parameter vector of the distribution $f_T(\cdot)$. Here $\theta = (\pi, \theta_2)$. The distribution of Y has a probability mass at zero and a positive continuous, right-skewed component, a form that was suggested by Cragg (1971). We refer to distributions having the form (2) as zero-adjusted distributions. Candidate distributions for $f_T(y; \theta_2)$ within the exponential family are the gamma and inverse Gaussian distributions, giving rise to the zero-adjusted gamma and zero-adjusted inverse Gaussian (ZAIG) distributions, respectively (Heller *et al.* 2006). In practice a more flexible distribution may be needed for T , e.g. the BCPE or BCT distributions (Rigby and Stasinopoulos 2004, 2006), or the skew-normal distribution (Chai and Bailey 2008).

Incorporation of covariates Following Rigby and Stasinopoulos (2005), who specify generalized additive models for the location, scale and shape parameters of a variety of distributions, we specify the following models on π and each of the parameters in θ_2 , denoting each parameter generically as ψ :

$$g(\psi) = x' \beta + \sum_{j=1}^J s_j(w_j),$$

where x and w_j for $j = 1, 2, \dots, J$ are covariate vectors for ψ , which may be overlapping or distinct; s_j for $j = 1, 2, \dots, J$ is a smooth nonparametric function, typically a smoothing spline; and g is an appropriately chosen link function for ψ . Usually the logarithmic link is used for μ and other non-negative parameters.

4 Estimation

Zero-adjusted models are implemented in the R package `gam1ss` (Stasinopoulos *et al.* 2006). Maximum (penalized) likelihood estimation is used to estimate the parameters β and smoothing functions s_j . The penalized log likelihood function is maximized iteratively using either the RS or CG algorithm of Rigby and Stasinopoulos (2005), which in turn uses a back-fitting algorithm to perform each step of the Fisher scoring procedure. Both RS and CG algorithms use the log likelihood of the data, and its first derivatives (and optionally expected second derivatives) with respect

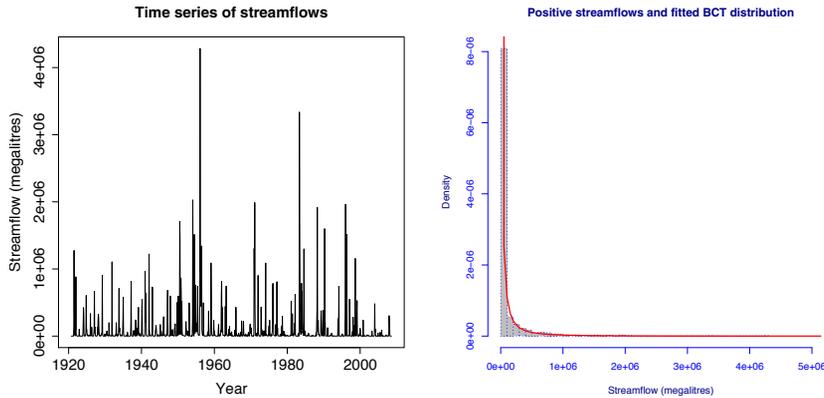


FIGURE 1. Monthly streamflows, Balonne River, 1920–2008

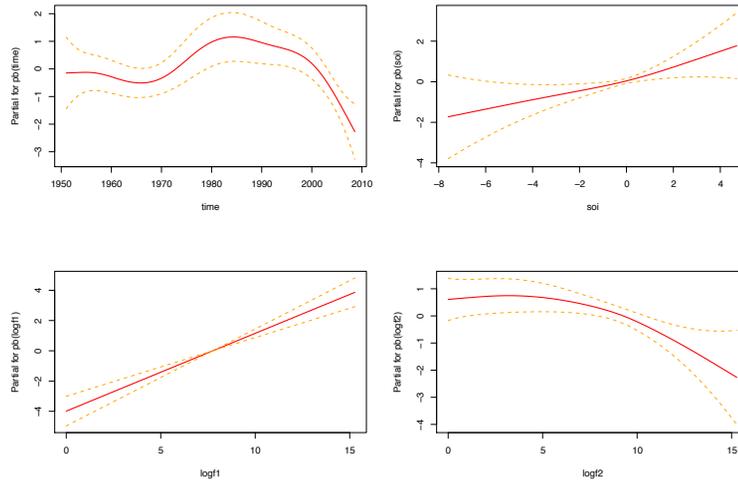
to distributional parameters. The CG algorithm, a generalization of the algorithm used by Cole and Green (1992), additionally uses the expected cross derivatives. The nonparametric smoothing functions s_j are penalized splines, i.e. P-splines (Eilers and Marx 1996). The smoothing effective degrees of freedom for the P-splines are chosen automatically using internal maximum likelihood (treating the second order difference of the P-spline basis functions coefficients as random effects), Pawitan (2001).

5 Monthly streamflows of Balonne River

Monthly streamflow data for the Balonne River, Queensland, Australia are available from 1920 to 2008, and are shown in the left panel of Figure 1. Modelling streamflow in this semi-arid region is of hydrological, ecological and economic interest. Eleven percent of the streamflows are zero, and the positive streamflows are strongly right-skewed. A histogram of these, with fitted BCT distribution, is shown in the right panel of Figure 1. While this is a fit of the marginal distribution only, the good fit is indicative that the BCT could be appropriate as the response distribution for T in the regression model.

As the monthly total streamflow (SF) is observed, but not the individual flow episodes, these data fall under Case 1. We adopt the approach of Model 1.2, and use the BCT distribution for T .

SOI (Southern Oscillation Index - monthly air pressure difference between Darwin and Tahiti) and sea surface temperatures are indicators of ENSO (El Niño Southern Oscillation) (Trenberth 1997). Rainfall patterns across eastern and north eastern Australia are influenced by ENSO (Ropelewski

FIGURE 2. The fitted smoothing functions for the π model

and Halpert 1987). Hence it is expected that SOI and sea surface temperatures will be predictive of rainfall and therefore streamflow. Time; month; SOI; and relevant indexes of sea surface surface temperatures EOF1 and EOF2 (first and second empirical orthogonal functions of sea surface temperature); and NINO3 (average sea surface temperatures in the region $90^\circ\text{W} - 150^\circ\text{W}$ and $5^\circ\text{S} - 5^\circ\text{N}$) are considered. Models are fitted to data from 1951 onwards, as the covariates are available from this time. Parameters to be fitted are π and the parameters of the BCT distribution: $\theta_2 = (\mu, \sigma, \nu, \tau)$. Model selection was performed using the SBC. Interactions between smooth terms were checked for by fitting smooth surfaces.

5.1 Model for π

The following model was selected for π :

$$\ln \frac{\hat{\pi}}{1 - \hat{\pi}} = 5.32 + s(t) + s(\text{SOI}) + s(F_{t-1}) + s(F_{t-2}) + 1.03 y_1$$

where t is calendar time, $F_{t-i} = \log(SF_{t-i} + 1)$ are lag values of the logarithm of the streamflow (adding 1) and $y_1 = I(y_{t-1} = 0)$ is a binary indicator variable, indicating whether the first lag of streamflow is zero or not. The function $s()$ is a P-spline. The effects of the smoothed covariates on the log-odds of a positive streamflow are shown in Figure 2. It is interesting to see whether these correspond to actual events in the historic records or expected catchment behaviour. In South Western Queensland there are two major drought periods, starting in the mid 1960's and late 2000. These

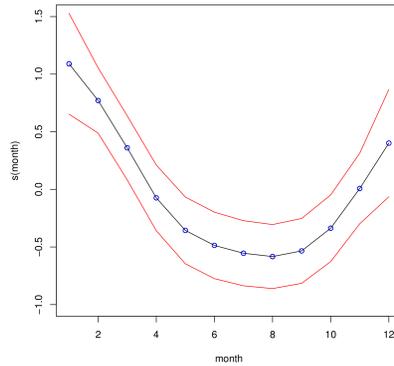


FIGURE 3. The fitted smoothing function for month for the μ model

two periods are well represented by the time covariate, which shows a decrease in probability of flow during these periods (top left panel Figure 2). The effect of SOI shows an almost linear positively sloping curve passing through zero (top right panel Figure 2). This conforms with the known positive relationship between SOI and rainfall/streamflow. Past streamflows are represented in the model by the log lagged flow covariates, and these hence summarize atmospheric and catchment processes. Both lower plots indicate that the memory in the system is significant for at least two months.

5.2 Model for θ_2

Two approaches were followed for modelling the positive streamflows: (1) modelling the log-transformed streamflows with a suitable distribution; and (2) modelling the raw streamflows, using a long-tailed distribution. The second approach, using the BCT distribution, produced a better-fitting model. This distribution has parameters μ , σ , ν and τ , where μ is the median, σ is approximately the coefficient of variation, ν is the skewness and τ is the kurtosis. The following models for μ and σ were selected:

$$\begin{aligned} \ln \hat{\mu} &= 64.2 - 0.03t + s(\text{month}) + 0.27 \text{SOI} + 0.19 \text{EOF2} \\ &\quad + 0.51 F_{t-1} + 4.76 y_1 \\ \ln \hat{\sigma} &= -10.1 + 0.005t + 0.09 \text{NINO3} + 0.06 \text{SOI} \\ &\quad - 0.07 F_{t-1} - 0.02 F_{t-2} - 0.70 y_1 \end{aligned}$$

Parameters ν and τ were estimated as $\hat{\nu} = 0.045$, $\hat{\tau} = 7.36$. The effect of month on $\ln \hat{\mu}$ is shown in Figure 3. The model preserves the seasonality in the positive streamflow, which is summer dominated in South Western

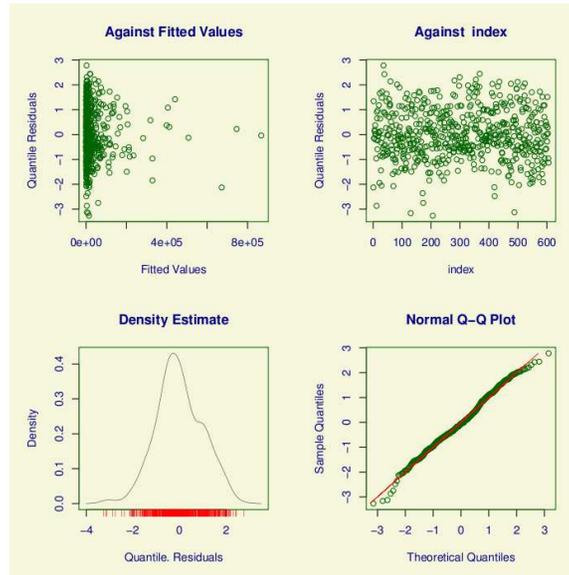


FIGURE 4. The residuals of the BCT fit

Queensland. The other covariates are adequately modelled by linear relationships. Residual plots for the positive part of the model are shown in Figure 4.

6 Conclusion

The method we present here is a sub-class of a wider family of models. In the hydrological application that we present, it produces a credible model which has potential usefulness in ecological prediction. The method is highly flexible in that any continuous right-skewed distribution may be used for the positive part of the model.

References

- Chai, H. S. and K. R. Bailey (2008). Use of log-skew-normal distribution in analysis of continuous data with a discrete component at zero. *Statistics in Medicine*, 27(18), 3643- 55.
- Cole, T. and P. Green (1992). Smoothing reference centile curves: The LMS method and penalized likelihood. *Statistics in Medicine*, 11, 1305-1319.
- Cragg, J. (1971). Some Statistical Models for Limited Dependent Variables with Applications to the Demand for Consumer Durables. *Econometrica*, 39(5), 829-844.

- Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B-splines and penalties, *Statistical Science*, 11(2): 89-121.
- Heller, G. Z., D. M. Stasinopoulos and R. A. Rigby (2006). The zero-adjusted inverse Gaussian distribution as a model for insurance data. In *Proceedings of the 21st International Workshop on Statistical Modelling*, 226-233.
- Jørgensen, B. and M. de Souza (1994). Fitting Tweedie's compound Poisson model to insurance claims data. *Scandinavian Actuarial Journal*, 69-93.
- Pawitan, Y. (2001). *In All Likelihood: Statistical modeling and inference using likelihood*, Oxford University Press.
- Rigby, R. A. and D. M. Stasinopoulos (2004). Smooth centile curves for skew and kurtotic data modelled using the Box-Cox Power Exponential distribution. *Statistics in Medicine*, **23**, 3053-3076.
- Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape. *Applied Statistics*, 54(3), 507-554.
- Rigby, R. A. and D. M. Stasinopoulos (2006). Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis. *Statistical Modelling*, **6**, 209-229.
- Ropelewski C. F. and M. S. Halpert (1987). Global and Regional Scale Precipitation Patterns Associated with the El Niño/Southern Oscillation. *Monthly Weather Review*, **115**, 1606-1626.
- Smyth, G. K. and B. Jørgensen (2002). Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling. *ASTIN Bulletin*, **32**, 143-157.
- Stasinopoulos D. M., R. A. Rigby and C. Akantziliotou (2006). gamlss: A collection of functions to fit Generalized Additive Models for Location Scale and Shape, R package version 2.0-0, <http://www.londonmet.ac.uk/gamlss/>.
- Stuart, A. and K. Ord (1994). *Kendall's Advanced Theory of Statistics. Vol. 1: Distribution Theory* (Sixth ed.). Edward Arnold.
- Trenberth, K. E. (1997). The Definition of El Niño. *Bulletin of the American Meteorological Society*, **78**, 2771-2777.

Improving the Calculation of Fix-Rate Bias in Automated Telemetry Systems

Miriam Hodge¹, Jennifer Brown¹ and Marco Reale¹

¹ Mathematics and Statistics Department, University of Canterbury, Private Bag 4800, Christchurch, New Zealand

Abstract: GPS and other radio tracking equipment are becoming more widely used by researchers for modelling animal habitat. In a typical monitoring program an animal will be fitted with a tracking collar. This tracking collar will fix the animal's location at a set time interval. These fixes of the animal's location can then be cross referenced on a digital map (GIS) containing habitat information and the animal's preferred habitat can be modelled.

Care must be used in modelling the habitat because radio tracking collars have different transmission probabilities in different habitats. The habitat observations are biased towards habitats that allow good transmission. One way to minimise this bias is to weight observations by a measure of transmission quality.

Researchers have attempted to estimate the detection weighting by placing stationary collars in the study area and recording the fix-rate. The results of these studies are unsatisfactory because stationary collars do not account for animal movement and behaviour. Johnson (1998) used a surrogate for stationary collars by analysing 6 hour time periods where the animal was relatively stationary. We will develop this method further by incorporating the non-stationary sites in the detection rate calculation.

Keywords: fix-rate; bias; telemetry collar.

1 Source of Fix-Rate Bias

The use of radio collars to collect information on habitat can be biased when habitat influences detection. (Lewis 2007, D'Eon 2002) The number of fixes for an animal at a location, f , is the number of times a fix is attempted from that location. These fixes are either successful, f_s , or unsuccessful, f_u .

$$f = f_s + f_u$$

Unsuccessful fixes can occur for many reasons. Links have been made between fix-rate and many habitat qualities to include slope (Lewis, 2007), tree density (Rumble Lindzey 1999), and terrain conditions (D'Eon 2002). Fix-rate, r , at a location is the number of successful radio fixes of an animal in that location divided by the total number of attempted radio fixes.

$$r = \frac{f_s}{f_s + f_u}$$

The detection weighting, w , at a given location is the inverse of the detection rate. Applying the detection weighting corrects for fix-rate bias.

$$w \cdot f_s = \frac{f_s + f_u}{f_s} \cdot f_s$$

$$w \cdot f_s = f$$

One technique to estimate fix-rate bias is to place collars in different locations within the study area and recording the fix-rate. Because each collar is stationary at a location the fix-rate at that location is the number of successful fixes divided by the number of fix attempts. Placing collars in each habitat type in a large study area very expensive and time consuming. Additionally, fix-rates from stationary collar studies do not accurately predict fix-rates in collars worn by animals. This disparity has been linked to animal movement and behaviour (Edenius 1997, Moen 1996).

Johnson (1998) used a surrogate for stationary collars by identifying time periods when the animal was relatively stationary and measuring the fix-rate in these periods. To find stationary periods they broke the day into four time periods (TP) each lasting 6 hours. Because the actual location of the animal is unknown, Johnson used the arithmetic mean of the successful fixes during a time period as the location. The animal was judged stationary if it had at least four fixes in the time period and one or less fixes were more than 200 metres from the mean location of the observed animal fixes. He recorded the detection weighting at each stationary site as the total number of fix attempts divided by the number of successful fixes in the time period. Because there are not stationary time periods in every location, Johnson goes on to calculate the detection weighting for all locations in the forest by employing five linear models of the form, $\mathbf{Y} = \mathbf{X} \beta + \epsilon$, where \mathbf{Y} is the vector of observation rates, \mathbf{X} is a design matrix of environmental variables, β is the design matrix and ϵ is the error term. Johnson's most successful model, the kriging model on a 180 m pixel grids omitted the environmental covariates.

Instead of using a linear model, we will estimate the weighting function at every site by assuming that at each location we are doing repeated samples and that over the length of the study the probability of radio transmission at a given location is constant and that the amount of time the animal spends in each location during a time period is constant. We then apply MacKenzie's (2002) detection rate technique to define the probability of detecting each fix and estimate the weighting function with Markov Chain Monte-Carlo techniques.

2 Calculation of Fix-Rate Bias

MacKenzie (2002) developed a technique for incorporating detection probabilities for animals into large-scale site occupancy surveys. We adjust this

technique by detecting locations instead of animals to incorporate fix-rate bias in radio telemetry studies. MacKenzie builds his model on two assumptions. The first assumption is that the area of inference is too large to be surveyed. The second is that detectability is not perfect. MacKenzie's assumptions are met by radio telemetry studies. The area in use by all animals is too large to be surveyed and the fix-rate is not one-hundred percent.

We define presence at a location as a six hour time period, TP, during which an animal has one successful fix at that location. An animal can be present at multiple locations during the same time period if the animal has successful fixes at multiple locations during a time period. We construct our detection probabilities by looking at fix records one time period at a time. In a given time period $p_{(p,1)}$ is the probability of being at location 1, $1-p_{(p,1)}$ is the probability of not being at location 1, $p_{(d,1)}$ is the probability of being detected at location 1, and $1-p_{(d,1)}$ is the probability of not being detected at location 1. For example, the probability of the following detection history at location 1

$$f_1, f_1, f_u, f_u, f_u, f_2, f_1$$

where f_1 is a fix at location 1, f_2 is a fix at location 2, and f_u is an unsuccessful fix. Is as follows:

$$(p_{(p,1)}p_{(d,1)})^3((1-p_{(p,1)})p_{(d,2)})(p_{(p,1)}(1-p_{(d,1)})+(1-p_{(p,1)})(1-p_{(d,?)})^3$$

assuming that all location visited by the animal in the time period have the same detection probability, this simplifies to

$$(p_{(p,1)}p_d)^3((1-p_{(p,1)})p_d)(1-p_d)^3$$

3 Results for Sample Data Set

The sample data set is the Starkey Experimental Forest and Range in Oregon, USA. The same data set from used by Johnson (1998). The 10,102-ha reserve is surrounded by a game-proof fence and includes diverse topography. Elk (*Cervus elaphus*), mule deer (*Odocoileus hemionus*), and cattle were fitted with radio collars and tracked by an Automated Telemetry System (ATS). For more detailed information on the ATS, the tracking data, and habitat information see the U.S. Forest Service web site, <http://www.fs.fed.us/pnw/starkey/>.

The survey area is broken into a grid with each pixel being 30 m to a side. In the MCMC model a location is defined as a single pixel in this grid. Each location, L, has a distinct probability of the animal being at the site, $p_{(p,L)}$. The probability of detection p_d is assumed constant over the study area. The model parameters are estimated with Markov Chain Monte Carlo

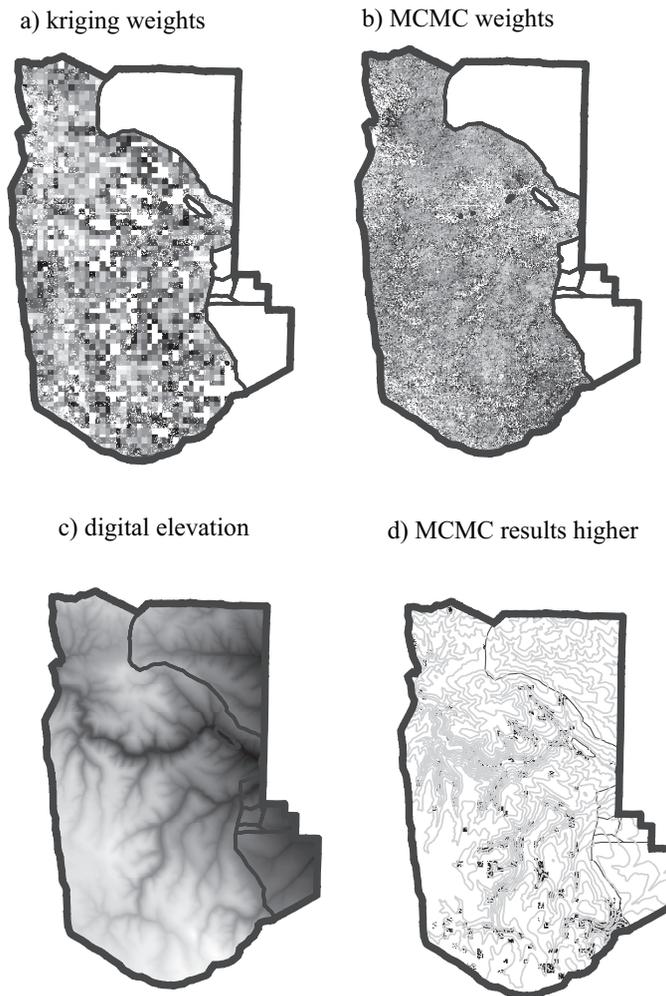


FIGURE 1. Models applied to Starkey data. Digital map boundaries, elevation contours, digital elevation model, and Johnson's weightings were obtained from U.S. forest service web site. The grey scale in map a and b is scaled from low to high fix-rate with dark to light shades.

using Winbugs software. (Lunn 2000) The $p_{(p,L)}$ values are given a vague priors.

The MCMC model has greater variation and higher average weighting than

Johnson's kriging model (1998). Figure Maps a), b) and c) show that both kriging model and the the MCMC model follow the digital elevation model of the study area. Figure Maps d) shows the elevation contours in grey and the locations where the MCMC model gives a higher weighting than the kriging model in black. The MCMC model is showing more missing fixes in areas with low elevation adjacent to steep slopes.

4 Future Work

A metric which can quantify the differences between models would facilitate the comparison of models developed with different theoretical frameworks. A possible source for such a metric are the deformation metrics being developed in anatomical fields to compare highly irregular body parts such as portions of the brain.

In addition to comparison metrics, we are looking to incorporate habitat and animal covariates into our model of detection probability, p_d . We will introduce a hierarchical model of p_d . This will allow for different behaviours by different animals in the same location.

We are also looking at incorporating adjacency information into the model of presence, p_p , using Geobugs component of the Winbugs software.(Lunn 2000) The MCMC probability of an animal being present at a location is calculated independently of all of the other location values. Incorporating adjacency information will allow us to recover the geographic relationship between locations that Johnson (1998) leverages through kriging.

5 Conclusion

Johnson (1998) considered one source of variation due to radio transmission and carefully constructed his stationary time periods to eliminate all other sources of variation. We introduce a second source of variation, namely detectability, that allows us to look at time periods that Johnson (1998) discarded from his model fitting process. We are able to fit our model based on all time periods where the automated telemetry system is working, not just those where the animal was stationary.

References

- D'Eon, Robert G., Serrouyn, Robert, Smith, Graham, and Kochanny, Christopher O., (2002). GPS radiotelemetry and bias in mountain terrain, *Wildlife Society Bulletin*, **30**, 430-439.
- Edenius, L, (1997). Field test of a GPS location system for moose Alces alces under Scandinavian boreal conditions, *Wildlife Biology*, **3**, 39-43.

- Johnson, Bruce K., (1998). Migrating Spatial Differences in Observation Rate of Automated Telemetry Systems, *The Journal of Wildlife Management*, **62**, 958-967.
- Lewis, Jesse S., Rachlow, Janet L., Garton, Edward O., Vierling, Lee A., (2007). Effects of habitat on GPS collar performance: using data screening to reduce location error, *Journal of Applied Ecology*, **44**, 663-671.
- Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D., (2000) WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**, 325–337.
- MacKenzie, Darryl I., Nichols, James D., Lachman, Gideon B., Droege, Sam, Royle, J. Andrew and Langtimm, Catherine A.,(2002). Estimating Site Occupancy Rates When Detection Probabilities Are Less Than One, *Ecology* , **83**, 2248-2255.
- Moen, Ron, Pastor, John, Cohen, Yosef and Schwartz, Charles C., (1996). Effects of Moose Movement and Habitat Use on GPS Collar Performance, *The Journal of Wildlife Management*, **60** 659-668.
- Rumble, M.A. and Lindzey, F., (1997). Effects of forest vegetation and topography on global positioning system collars for elk. *Resource Technology Institute Symposium*, **4**, 492501.

Bayesian Modelling of Grainsize Distributions

S. Huzurbazar¹ and Jarrett J. Barber ¹

¹ Department of Statistics, University of Wyoming Dept 3332, 1000 E. University Avenue, Laramie, WY 82071, USA

Keywords: latent modelling, log-hyperbolic; MCMC.

1 Introduction

The analysis of grain size or particle size data is a problem dating back more than a century, with most of the literature concentrated in the geological sciences. The first modern statistical approach to the problem started with the work of Barndorff-Nielsen (1977), in which he derived the hyperbolic family of distributions using characteristics of empirical distributions of particle sizes of wind blown sand as documented extensively by R.A. Bagnold. Since the natural logarithm of the particle size data is modelled, the distribution is called the log-hyperbolic. Work by Bagnold and Barndorff-Nielsen (1980), Barndorff-Nielsen et al (1982) and Fieller et al. (1992), among others, provided more applications of the log-hyperbolic to grain sizes in different environments. However, use of the distribution has been limited in the geological sciences, partly due to the strong tradition of using the log-normal and partly due to the computational difficulties with fitting the log-hyperbolic. We approach the problem from a Bayesian perspective with the goal of obtaining posterior distributions for the parameters of the log-hyperbolic distributions. We eventually hope to be able to model various issues arising in data collection.

The problem motivating the present work stems from interactions between the first author and Elizabeth Hajek and Dr. Paul Heller, sedimentologists at the University of Wyoming. The data used in this study were collected by Dr. R. Lynds (2005) with the goal of studying fluvial sediments in modern and ancient systems, in an effort to characterize sediment transport in ancient systems. As part of a longer project, the first step is modelling of grain size distributions using the data collected by Lynds.

1.1 Data Description

Data were collected in Nebraska, USA, from 3 modern rivers, the Calamus, the Northloup and the Niobrara. Three different sediment types, bedload,

slackwater and suspended, were sampled from each of the above rivers. Data were also collected in some ancient rivers from the Kayenta formation in S.W. Colorado and Utah. Each of the samples was processed for sediment size in the Sedimentary Geology Laboratory at the Massachusetts Institute of Technology. The fine-grained (< 0.09 mm) fraction was analyzed with a Horiba LA-300 laser particle-size analyzer (LPSA), and the coarse-grained material (> 0.09 mm) was analyzed using a Retsch Technology digital image-processing particle-size analyzer (CAMSIZER). The LPSA uses a diode laser to measure grain sizes from 0.001-0.1 mm in diameter and the CAMSIZER uses digital photographic images to measure grain sizes ranging from 0.05-30 mm in diameter.

The final measurements available to us are the weights of each sample within each particle size category resulting in grouped data. Note that the count of the particles within each size category is unavailable. These measurement techniques are similar to those described in Fieller et al.(1992), except their samples were obtained via sieving. The particle size classes for our data vary from $0.877 \mu\text{m}$ at the lower limit to $7210 \mu\text{m}$ at the upper limit. The final grain-size data are a compilation of the data from the fine- and coarse-grained components.

2 Frequentist Analysis

The log-hyperbolic distribution has the following form:

$$g(x|\pi, \zeta, \delta, \mu) = \frac{\exp\{-\zeta[\sqrt{1+\pi^2} \sqrt{1+(\frac{x-\mu}{\delta})^2} - \pi(\frac{x-\mu}{\delta})]\}}{\{2\delta\sqrt{1+\pi^2} K_1(\zeta)\}} \quad (1)$$

where $x \in R$ is the log-grain size, $K_1(\cdot)$ is the modified Bessel fct of 3^{rd} kind, $\mu \in R$, $\delta > 0$ are the location and scale parameters with $\pi \in R$ and $\zeta > 0$ capturing asymmetry and peakedness. The limiting distributions are the log-normal and the log-skew-Laplace.

In grain size settings, data are reported by size intervals, $[c_{i-1}, c_i]$, and we are led to consider truncated and renormalized versions of (1):

$$f_i(x|\theta) = \frac{g(x|\theta)}{p_i} \quad c_i \leq x \leq c_{i+1}, \quad i = 1, \dots, k, \quad (2)$$

where

$$p_i = \int_{c_{i-1}}^{c_i} g(x|\theta) dx, \quad (3)$$

and $\theta \equiv (\pi, \zeta, \delta, \mu)^T$.

A further complication arises since the total weights and not the counts of the grains in each interval are available. Assuming that the size (diameter) of a grain is proportional its weight, and that the proportionality constant

TABLE 1. Summary of fitted distributions by sediment type.

Sediment	Samples	log-hyperbolic	Limiting cases	2 comp mixture
Suspended				
modern	39	32 (82%)	6 log normal	1 (2.5%)
ancient	36	0 (0%)	0	36 (100%)
Slackwater				
modern	27	2 (7.4 %)	1 log-S-L	24 (88%)
ancient	46	0 (0%)	0	46 (100%)
Bedload				
modern	67	48 (72%)	14 log-S-L	4 (6 %)
ancient	93	0 (0%)	0	93(100%)

is the same across all grains, (1) is referred to as the mass-size model as opposed to the size model when considering individual grains. The parameters of the mass-size and size distributions are linearly related for a reparameterized version of (1). Working with the (normalized) grouped weights as data, a likelihood-like ‘likeness function’ (Barndorff-Nielsen, 1977) is used to estimate parameters in the mass-size model.

The first step in our analysis has been to analyze each sample from each sediment type from each river using this methodology as implemented in the program Shefsize available from Robson, Fieller and Stillman (1997). The program is able to fit unimodal log-hyperbolic, log-normal and log-skew-Laplace distributions, along with a 2 component mixture of the latter. A modified χ^2 type statistic is used to pick the best fit. We refer to Table 1 for a summary of the best fitted distributions within these limitations. In modern rivers, for unimodal distributions, the log-hyperbolic appears to fit most often, and for some sediment types and for all ancient rivers mixtures fit best.

A major limitation of the log-hyperbolic model fitting appears to be non-convergence of the algorithm that maximizes the likeness function (p.138, Fieller et al, 1992); a situation that appears to occur frequently in the presence of measurement error. Another limitation is the inability to model data across different samples, rivers and sediment types. Analyses of samples using such frequentist methods are done sample by sample. We suspect that Bayesian methodology can help overcome some of these problems. As a start, we approach Bayesian modelling for data from one sample, which can later be expanded to include other samples, sediments and rivers.

3 Bayesian Modelling

Here, we exploit the idea of a latent (unobserved) grain size, and we concentrate on the size distribution, not the mass-size distribution.

3.1 A Latent Grain Size Specification

Consider the (log) grain sizes, x_{ij} , $i = 1, \dots, k$, $j = 1, \dots, n_i$, where n_i is the (latent) number of grains in each group, and $N \equiv \sum_{i=1}^k n_i$ is the total number of grains in groups. Grains commonly are assumed to be spherical so that the mass of a grain with diameter $\exp(x)$ is

$$m(x) = (\text{Density}) \frac{4}{3} \pi \left(\frac{\exp(x)}{2} \right)^3. \quad (4)$$

We do observe the masses, w_i , of grains in (log) size classes $[c_{i-1}, c_i]$, $i = 1, \dots, k$. Adding independent measurement errors, $\varepsilon_i \sim N(0, \sigma^2)$, to the observed masses, we obtain $w_i = \sum_{j=1}^{n_i} m(x_{ij}) + \varepsilon_i$ as conditionally independent normal variates,

$$w_i \mid \{x_{ij}\}, \sigma \sim N \left(\sum_{j=1}^{n_i} m(x_{ij}), \sigma^2 \right). \quad (5)$$

Since we do not know n_i , we give it a stochastic specification to account for its uncertainty. If we know θ and $N = \sum_{i=1}^k n_i$, then $n_i \sim \text{binom}(N, p_i)$. Since it is reasonable to assume large N for our application, $n_i \sim \text{Pois}(Np_i)$, approximately. Of course, we do not know N , but our development so far suggests the model $n_i \sim [n_i] \equiv \text{Pois}(N^*p_i)$, where we choose to model N^* as $[N^*] \equiv \text{gamma}(\alpha, \beta)$. Note N and N^* coincide conceptually only when we use all classes for which we have weight measurements, else we expect $N < N^*$.

3.2 A Revised Specification

While the model above is a conceptually appealing representation of observed weights as sums of individual grain weights (plus error), it has computational drawbacks. In particular, using MCMC methods in a Bayesian framework, the full conditional distribution for the latent x_{ij} sizes is not of known form. Thus, we would have to seek an alternative to Gibbs sampling for this full conditional. This is not bad, per se, but undoubtedly, $\sum n_i$ will be very large, making sampling all x_{ij} problematic. Moreover, the dimensions of the vector with elements of grain size, x_{ij} , changes with n_i across sampling iterations, creating a trans-dimensional sampling problem (Green, 1995, Sisson, 2005). Since our primary interest lies with θ and not individual grain sizes, and since maintaining latent grain size does not seem to facilitate computation, we seek an alternative model that retains some properties of the above model but avoids difficulties of considering sizes explicitly.

We choose to marginalize over the x_{ij} in (5) to get

$$w_i | \theta, \sigma \sim [w_i] \equiv N(n_i \mu_i, \sigma^2 + n_i \sigma_i^2), \quad (6)$$

where μ_i is the expected mass of a grain in the i^{th} group,

$$\mu_i = \int_{c_{i-1}}^{c_i} m(x) f_i(x | \theta) dx, \quad (7)$$

and σ_i^2 is the corresponding variance

$$\sigma_i^2 = \int_{c_{i-1}}^{c_i} (m(x) - \mu_i)^2 f_i(x | \theta) dx. \quad (8)$$

This revision requires integration for p_i , μ_i , and σ_i^2 , but k is usually comparatively small, and these computations are relatively quick. Moreover, we do not have the convergence problems that are likely when maintaining a very large number of latent grain sizes.

We can write the full-probability model to which the posterior distribution is proportional,

$$[\theta, \{n_i\}, N^*, \sigma | \{w_i\}] \propto \prod_{i=1}^k [w_i][n_i][N^*][\theta][\sigma^2], \quad (9)$$

where $[\theta]$ and $[\sigma^2]$ are prior specifications. For convenience, we assume *a priori* independence among the components of θ , using disperse normal distributions for $[\mu]$ and $[\pi]$ and disperse exponential distributions for $[\delta]$ and $[\zeta]$. That is, $[\theta] = [\mu][\pi][\delta][\zeta]$. We use Jefferey's prior, $[\sigma^2] \propto 1/\sigma^2$.

4 Simulated Data, Sampling, Results and Discussion

To test our model, we simulated 1.5 million grains (log diameters) from (1) with parameter values $\delta = 0.522$, $\mu = 5.165$, $\pi = 0.143$ and $\zeta = 2.474$. Using lower and upper limits of $c_0 = 4.484$ and $c_k = 7.843$ (log μm scale), we created $k = 22$ classes, the first 12 having width 0.136 and the latter 10 width 0.173. The range of size values and parameter values correspond to observed data from some bedload samples from the Calamus river. We then used (5) to compute w_i "data" using density of quartz ($2.65\text{g}/\text{cm}^3$) and $\sigma^2 = (0.002)^2$. About $N = \sum n_i = 1.46$ million grains, totaling about $\sum w_i = 49.5$ grams, fell into the 22 classes, the remainder falling in the tails of g . Note N^* is analogous to 1.5 million, not 1.46 million.

The full-conditional distribution for θ is non-standard, and we use a Metropolis-Hastings (MH) step with a four-variate random walk normal proposal distribution centered at the current iteration's value for θ and oriented using

the empirical covariance matrix of θ computed from preliminary sampling runs. This block sampling works well to address the high correlation among the components of θ ; see Figures 1 and 2. The gamma prior for N^* is conditionally conjugate with the Poisson specifications for the n_i and results in a gamma full-conditional. The full-conditionals for the n_i are non-standard discrete distributions, and we used an MH step for each with a discrete “tent” or “triangle” proposal distribution centered at the current value of n_i . The full-conditional for σ^2 is non-standard, and we use an MH step with a random walk gamma proposal centered at the current iteration’s value of σ^2 . To explore sampling behaviour, we also introduced latent variables $m_i \sim N(n_i\mu_i, n_i\sigma_i^2)$, now giving $w_i \sim N(m_i, \sigma^2)$. This results in an inverse-gamma and normal full-conditionals for σ^2 and the m_i , respectively. Generally, sampling performance was comparable between the non-latent and latent implementations. Results shown here use the latent implementation and are based on the last 8,000 iterations of 20,000 iteration chain wherein proposals were tuned for the first 4000 iterations.

We used R: A language and environment for statistical computing to perform all computations (R Development Core Team, 2008).

Results are shown in Figures 1 and 2 and Table 2. We see that all marginal posterior summaries indicate that the corresponding “true” values are well within their respective 95% credible intervals. All marginals were unimodal and approximately symmetric. We omit summaries of the n_i for space.

Finally, note that Figures 1 and 2 result from using a small measurement error variance of $(0.002)^2$, and reveal the potential for non-identifiability in the form of high correlations between pairs of components of θ . A form of non-identifiability was observed when using larger values of σ , and we believe this is related to the high correlations observed in Figures 1 and 2. Given this, we believe measurement error modelling to be key to better model fitting with such data, and we will investigate different error models and reparameterizations of (1).

TABLE 2. Summary of Posterior distributions.

Parameter	2.5%	mean	50%	97.5%
δ	.5069941	.5233998	.5234937	.5383338
μ	5.148861	5.159868	5.159325	5.172596
π	.1332083	.1485737	.1493918	.1617763
ζ	2.360701	2.494168	2.495863	2.617763
σ	.001509549	.003529373	.003231754	.007228468
N^*	1.498794e+06	1.502732e+06	1.502630e+06	1.507430e+06

Acknowledgments: The first author acknowledges the Donors of the

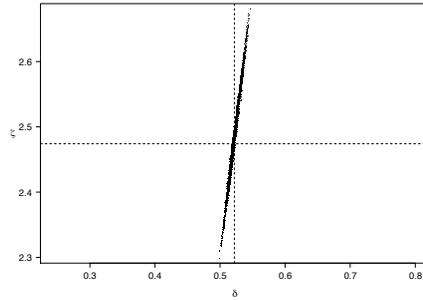


FIGURE 1. Posterior scatterplots for scale (δ) and peakedness (ζ) indicating high correlation and near non-identifiability. (Dashed lines indicate values used to generate data)

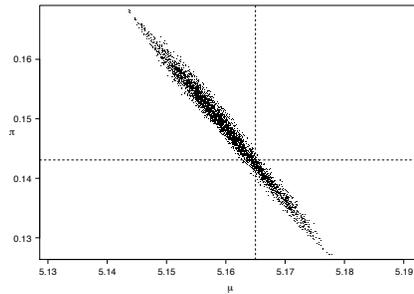


FIGURE 2. Posterior scatterplots for location (μ) and asymmetry (π) illustrating high correlation and near non-identifiability. (Dashed lines indicate values used to generate data)

American Chemical Society Petroleum Research Fund for partial support of this research. Special thanks to Elizabeth Hajek, Rainie Lynds, Paul Heller at the University of Wyoming, and David Mohrig at the University of Texas-Austin, for the original data.

References

- Bagnold, R.A. and Barndorff-Nielsen, O. (1980) The pattern of natural size distribution. *Sedimentology*, **27**, 199-207.
- Barndorff-Nielsen, O. (1977) Exponentially decreasing distributions for the

- logarithm of particle size. *Proc. Royal Society London A*, **353**, 401-419.
- Barndorff-Nielsen, O., Dalsgaard, K., Halgreen, C., Kuhlman, H., Moller, J.T., Schou, G. (1982) Variation in particle size distribution over a small dune. *Sedimentology*, 29,53-65.
- Fieller, N.R.J., Flenley, E.C., Olbricht, W. (1992) Statistics of particle size data. *Applied Statistics*, **41**, 127-146.
- Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination *Biometrika*, 82, 711-98
- Lynds, R. (2005) *Fine-grained sediment in modern and ancient sandy braided rivers*. PhD Dissertation, University of Wyoming.
- Robson, D., Fieller, N., and Stillman, E (1997). *Shefsize User's Manual*. University of Sheffield, U.K.
- R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Sisson, S. A. (2005). Transdimensional Markov chains: a decade of progress and future perspectives. *Journal of the Am. Statist. Assoc.*, 100, 1077-1089.

Eigenvalues Application in Robust outlier Detection

N.K. Jajo¹ and K.M. Matawie²

¹ Australian Defence Organization, Brindabella Circuit, BP33-4-76, Canberra, ACT 2600, Australia. Email: Nethal.Jajo@defence.gov.au

² School of Computing and Mathematics, University of Western Sydney, Locked Bag 1797, Penrith South, NSW 1797, Australia. Email: K.Matawie@uws.edu.au

Abstract: This paper proposes a new method to detect outliers in linear regression. The procedure based on the eigenstructure of an influential matrix which is defined to be the Jackknifed matrix using Least Trimmed Squares method to obtain robust residuals. Outlier detection criteria is introduced and its prominent effect is presented using two well know examples.

Keywords: outliers; Jackknife, masking, swamping

1 Introduction

In literature, many outlier detection procedures were introduced based on identifying single outlier or multiple outliers in linear regression. Among those who defecated their research for single outlier detection are Beckman and Cook (1983) and Chatterjee and Hadi (1986). The more outliers that are suspected in our data the more complicated the problem will become. Masking (outliers hiding each others) and swamping (make non-outliers look like outliers) are two common obstacles facing multiple outlier detection procedures. A consecutive procedure were introduced by Marasinghe (1985) and Kinifard and Swallow (1990) to identify a specify set of outliers, which their number was suggested in advance. To overcome the problem of masking, a resampling procedures based on high breakdown estimators were suggested by many researchers, among them are Rousseeuw and van Zomeren (1990) and Rousseeuw and Leroy (2003). Pena and Yohai (1995) presented the detection of subset of outliers using eigenvalues of the influence matrix. The construction of their influence matrix based on assumption that the masking effect is due to the presence in the sample of blocks of influential observations having similar or opposite effects. For each of the non-null eigenvalue of the influential matrix, obtain the eigenvector. The co-ordinates of the eigenvector will be arranged in decreasing order, then a positive ratio between the consecutive components will be calculated. If the ratio is large enough (larger than a k) then that observation will be declared as suspected outliers. The power of this procedure based on the

choice of k . Pena and Yohai (1995) suggested $k = 2.5$ without justification. For outlier detection, remove all the suspected outliers and then use t -test with Bonferroni inequality to determine the cut-off value for declaring outliers. The determination of the value of k and the cut-off value is critical and it has more serious consequences than the procedures itself.

Our new eigenvalue method developed in this paper has been applied to detect multiple outliers in number of well known real data sets. The proposed method is based on the largest eigenvalues of the ‘influential matrix’. This matrix is defined to be the Jackknifed matrix using Least Trimmed Squares (LTS) method to obtain robust residuals. LTS is a robust method that will be used in this paper for its prominent high breakdown estimators which are equivariant for linear transformations on the \mathbf{x}_i and is related to projection pursuit which means that we can use the hat matrix for LTS residuals’ studentized, Rousseeuw and Leroy (2003). The new method shows an appeal working in all the data sets and in comparing with other methods in literature.

Next Section of this paper will introduce the influential matrix and its eigenvalues including the steps to detect the outliers, two examples with the graphs will also be presented. A brief conclusion is given in Section 3.

2 Influential matrix and outlier detection

Let us consider the general linear regression model where $\mathbf{y} = (y_1, \dots, y_n)'$ are generated by

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \epsilon \tag{1}$$

assuming that there are n data points $(y_i, x_{i1}, x_{i2}, \dots, x_{in})$. The following notations will be used in the rest of the paper: $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$, \mathbf{X} is $n \times p$ matrix with rows $\mathbf{X}'_1, \dots, \mathbf{X}'_n$, $\mathbf{b} = (b_1, \dots, b_p)'$ is the vector of regression coefficients and $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$ is the vector of regression errors, where the ϵ_i is independent random variable with $N(0, \sigma^2)$ distribution. Let the vector $\mathbf{t}_i = r_i/s_{(i)}\sqrt{1-h_{ii}}$ be the LTS robust studentized residuals where $s^2_{(i)} = \sum_{i=1}^n r_i^2/(n-p)$ and h_i is the i th column of the hat matrix.

Based on Cook’s statistics $\mathbf{t}'_i \mathbf{t}_i / p\sigma^2$, we define the $n \times n$ influential matrix \mathbf{M} as:

$$\mathbf{M} = \frac{1}{ps^2} \mathbf{t}\mathbf{t}' \tag{2}$$

where $s^2 = \sum_{i=1}^n t_i^2 / (n-p)$.

We are interested in the largest eigenvalue of the Jackknife covariance matrix $\mathbf{M}_{(i)}$. This is an $(n-1) \times (n-1)$ influential matrix after deleting the i th observation. Denoting the largest eigenvalue of \mathbf{M} by λ and that of population matrix by Λ . Then we have

$$\lambda \stackrel{L}{\sim} N(\Lambda, 2\Lambda^2/n) \tag{3}$$

or

$$\frac{\sqrt{n}(\lambda - \Lambda)}{\sqrt{2\Lambda}} \stackrel{L}{\sim} N(0, 1)$$

2.1 The diagnose method

Consider the largest eigenvalue $\lambda_{(i)}$ of the influential matrix $\mathbf{M}_{(i)}$, $i = 1, 2, \dots, n$. Suppose $\Lambda_{(i)}$ is the corresponding population value of $\lambda_{(i)}$. It is clear that $\Lambda = \Lambda_{(i)}$ if the observation i is not an outlier and $\Lambda < \Lambda_{(i)}$ if the observation i is outlier.

Now consider the hypotheses:

$$H_0 : \Lambda = \Lambda_{(i)}, \quad \text{if the } i\text{th observation is not an outlier against}$$

$$H_1 : \Lambda > \Lambda_{(i)}, \quad \text{if the } i\text{th observation is an outlier.}$$

Since $\mathbf{M}_{(i)}$ is influential matrix with $(n - 1)$ observations, then for $i = 1, 2, \dots, n$, we have

$$\lambda_{(i)} \stackrel{L}{\sim} N\left(\Lambda_{(i)}, \frac{2\Lambda_{(i)}^2}{n-1}\right) \quad (4)$$

Assuming α as the probability of type I error of the above test, that is, $\alpha = P(\lambda - \lambda_{(i)} > 0 | H_0 \text{ is true})$. Where the distribution of $\lambda - \lambda_{(i)}$ is given by

$$\lambda - \lambda_{(i)} \stackrel{L}{\sim} N\left(0, 2\Lambda^2\left(\frac{1}{n} + \frac{1}{n-1}\right)\right). \quad (5)$$

Thus, we will reject H_0 if

$$\frac{\lambda - \lambda_{(i)}}{2\lambda[(2n-1)/(n(n-1))]^{1/2}} > Z_\alpha, \quad \text{where } i = 1, 2, \dots, n.$$

This test procedure is repeated for all n observation and outliers are declared at the end. The steps involved this eigenvalue method can be summarized as: Step 1: Compute the influential matrices \mathbf{M} and $\mathbf{M}_{(i)}$ for $i = 1, 2, \dots, n$. Obtain the largest eigenvalue of \mathbf{M} and $\mathbf{M}_{(i)}$ $i = 1, 2, \dots, n$., Step 2: Test the hypothesis $H_0 : \Lambda = \Lambda_{(i)}$, against $H_1 : \Lambda > \Lambda_{(i)}$ for $i = 1, 2, \dots, n$ to determine the outliers., Step 3: Compute the normalized eigenvalue

$$\lambda_{(i)}^* = \frac{(\lambda - \lambda_{(i)})}{\sum_{i=1}^n (\lambda - \lambda_{(i)})}, \quad \text{for } i = 1, 2, \dots, n.$$

Step 4: Plot the points $\{(i, \lambda_{(i)}^*), i = 1, 2, \dots, n\}$.

2.2 Examples

Dilemma data, Hocking and Pendleton (1983), contains 26 observations with 3 regressor variables and 3 outliers (observations 11, 17 and 18). Second example is the Salinity data, Ruppert and Carroll (1980), contains 28 observations, three regressor variables and certainly two outliers (observations 5 and 16).

Following are the graphs (Figure 1 and 2) for the above two examples identifying and showing the outliers very clearly as described in section 2.

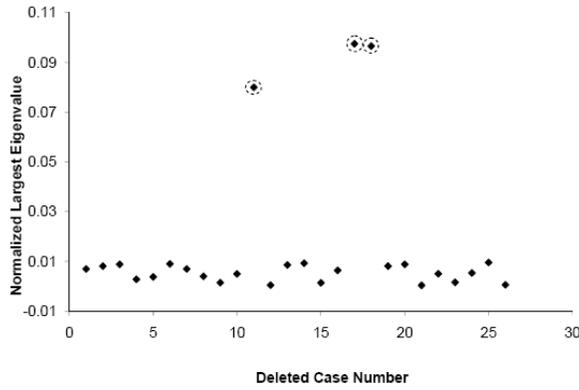


FIGURE 1. Jackknafied Largest Eignevalue for Dilemma data

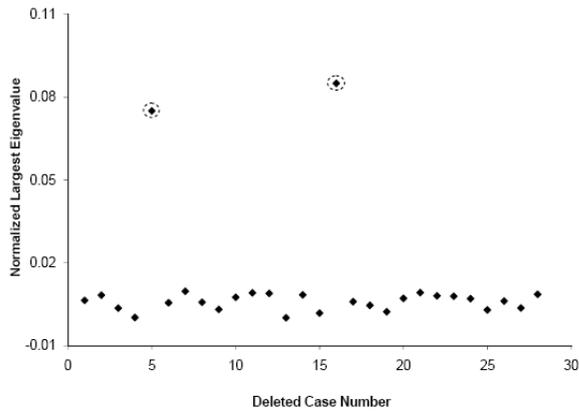


FIGURE 2. Jackknafied Largest Eignevalue for Salinity data

3 Conclusion

The method presented here showed the usefulness of the eigenvalues of the influential matrix to detect the outliers, and this is applicable when the data are less than 50% contaminated. The outliers, no matter in what order, can be identified and tested one by one rather than any subset identification, the distribution of the subset and the nominated cut-off value.

References

- D. Ruppert and R. J. Carroll (1980); Trimmed least squares estimation in the linear regression model. *J. Amer. Statist. Assoc.*, **75**: 828-838.
- D. Pena and V. J. Yohai (1995): The detection of influential subsets in linear regression using an influence matrix. *J. R. Statist. Soc. B*, 145-156.
- F. Kianifard and W. H. Swallow (1990): A Monte Carlo comparison of five procedures identifying outliers in linear regression. *Communications in statistics, Theory and Methods*, 91: 391-400.
- M. G. Marasinghe (1985): A multistage procedure for detecting several outliers in linear regression. *Technometrics*, 27: 395-399.
- P. J. Rousseeuw and A. M. Leroy (2003): *Robust Regression and Outlier Detection*. New York: Wiley
- P. J. Rousseeuw and B. C. van Zomeren (1990): Unmasking multivariate outliers and leverage Points. *J. American Stat. Assoc.*, 85, 633 -639.
- R. J. Beckman and R. D. Cook (1983): Outlier....s. *Technometrics*, 25, 119-163.
- R. R. Hocking and O. J. Pendleton (1983); The regression dilemma. *Communication in Statistics, part A- Theory and Methods*, **67**: 388-394.
- S. Chatterjee and A. S. Hadi (1986): Influential observations, high leverage points, and outliers in linear regression. *Statist. Sci.*, 415-416.

Prediction of binary response using multivariate longitudinal profiles: Study on chronic hepatitis B patients

Arnošt Komárek¹, Bettina E. Hansen²⁴, Harry L.A. Janssen⁴, Emmanuel Lesaffre²³

¹ Dept. of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University in Prague, Sokolovská 83, 186 75 Praha 8, the Czech Republic. E-mail: Arnost.Komarek@mff.cuni.cz

² Dept. of Biostatistics, Erasmus University Rotterdam, Dr. Molewaterplein 50, 3015 GE Rotterdam, the Netherlands.

³ Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Katholieke Universiteit Leuven, Kapucijnenvoer 35, Blok D, Bus 7001, 3000 Leuven, Belgium and Universiteit Hasselt, Belgium.

⁴ Dept. of Gastroenterology and Hepatology, Erasmus Medical Center Rotterdam, Dr. Molewaterplein 40, 3015 GD Rotterdam, the Netherlands.

Keywords: Conditional prediction; Linear mixed model; Marginal prediction; Random effect prediction.

1 Introduction

Nowadays several treatment options are available for patients with chronic hepatitis B. However, the hepatitis B virus remains quite difficult to eliminate and only 2-36% of the treated patients are cured. Peg-interferon (PEG-INF) has proven effective but also has its limitations regarding multiple and possible serious side-effects. Hence, it is important to predict as early as possible whether the patient will respond positively to the treatment. During therapy the patient is monitored at frequently scheduled follow-up visits and several markers are measured to anticipate continuation of therapy. In this contribution, we developed and compared several prediction models based on the past history of the patient for the treatment response (*cured/not cured*).

2 Clinical Study

In the international HBV9901-trial on chronic hepatitis B, 226 HBeAg positive patients were randomized to receive either PEG-INF mono-therapy for 52 weeks or PEG-INF in combination with lamivudine. The primary outcome, defined as HBeAg-negative at 26 weeks post-treatment, was achieved

in 36%. Earlier reports of the trial (Janssen et al., 2005) showed a higher probability for this response among patients with hepatitis B virus (HBV) genotype A, low baseline viral load, high baseline ALT (measure of disease activity) and previous treatment. Only the subset of patients ($n = 125$) with PEG-INF mono-therapy and HBV genotype A, B, C, and D is studied here. During treatment the viral load, HBV DNA (copies/ml), the disease activity and ALT were measured every 4th week until the end of follow-up (week 78) and their logarithms will be used as predictors (markers) for the overall response.

3 Methods

Let us first introduce some notation. For the i th subject ($i = 1, \dots, N$), let $\mathbf{Y}_{i,j} = (Y_{i,j,1}, Y_{i,j,2})'$ denote a random vector of \log_{10} (viral load) and \log (ALT) obtained at visit j ($j = 1, \dots, n_i$). Let R_i be the dichotomized overall response of patient i (1 if HBeAg negative at week 78, 0 if HBeAg positive at week 78). Let \mathcal{I}_0 be a set of indices of patients with $R_i = 0$ and analogously, let \mathcal{I}_1 be a set of indices of patients with $R_i = 1$. Further, let $\mathbf{Y}_i = (\mathbf{Y}'_{i,1}, \dots, \mathbf{Y}'_{i,n_i})'$ be a random vector of markers obtained for the i th subject. Firstly, the multivariate linear mixed model (see, e.g., Morrell et al., 2005)

$$\mathbf{Y}_i = \mathbf{X}_i^g \boldsymbol{\beta}^g + \mathbf{Z}_i^g \mathbf{b}_i^g + \boldsymbol{\varepsilon}_i^g, \quad i \in \mathcal{I}_g \quad (1)$$

is fitted separately for patients from the \mathcal{I}_0 and \mathcal{I}_1 groups. In model (1), $\boldsymbol{\beta}^g$ is the fixed effects vector in the g -th overall response group ($g = 0, 1$). Further, \mathbf{b}_i^g is the random effect vector assumed to follow a normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{D}^g)$ and $\boldsymbol{\varepsilon}_i^g$ is the vector of random errors, independent of \mathbf{b}_i^g and also normally distributed according to $\mathcal{N}(\mathbf{0}, \sigma_g^2 I_{n_i})$. Finally, \mathbf{X}_i^g and \mathbf{Z}_i^g are design matrices for the fixed and random effects in the g -th group based on time, gender, age, virus genotype, previous treatment, To represent the prediction model as a flexible function of the evolution of the markers we implemented B-splines (Dierckx, 1993) for both fixed and random effects.

Let $\hat{\boldsymbol{\beta}}^g, \hat{\mathbf{D}}^g, (\hat{\sigma}^g)^2$, ($g = 0, 1$) be (RE)ML estimates of parameters of model (1) based on subset \mathcal{I}_g of the data. A new subject with observed history $\mathbf{Y}^{new} = (Y_{1,1}^{new}, Y_{1,2}^{new}, \dots, Y_{n,1}^{new}, Y_{n,2}^{new})'$ of \log_{10} (viral load) and \log (ALT) has an unknown value R^{new} . That is, it is unknown whether (s)he will respond positively to the treatment and our aim is to predict this response as early as possible with high certainty. Three approaches for prediction of $\pi_{1,new} = P(R^{new} = 1)$ based on estimates from the two ($g = 0, 1$) mixed models (1) (see, e.g., Morrell et al., 2007) will be compared. First, prior probabilities π_0, π_1 of the two response groups are given or estimated as proportions of HBeAg negative and positive patients in the data set. Then for a new subject, posterior probabilities $\pi_{0,new}, \pi_{1,new}$ are computed using

the Bayes rule as

$$\pi_{g,new} = \frac{\pi_g f_{g,new}}{\pi_0 f_{0,new} + \pi_1 f_{1,new}}, \quad g = 0, 1, \quad (2)$$

where $f_{0,new}$ and $f_{1,new}$ are quantities which depend on \mathbf{Y}^{new} and parameters estimated from the mixed models (1). Given an a priori chosen cut-off value for the probability, the new subject will be classified into the respective overall response group if either of $\pi_{0,new}$, $\pi_{1,new}$ exceeds the cut-off. Otherwise, the subject remains unclassified. In the remainder of the paper, let $\varphi(\cdot | \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the density of the (multivariate) normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Suppose further that $\mathbf{X}^{0,new}$, $\mathbf{Z}^{0,new}$ are design matrices for the mixed model (1) would R^{new} be equal to 0. Similarly, let $\mathbf{X}^{1,new}$, $\mathbf{Z}^{1,new}$ be design matrices for the mixed model (1) would R^{new} be equal to 1. Note that $\mathbf{X}^{0,new}$, $\mathbf{X}^{1,new}$ and $\mathbf{Z}^{0,new}$, $\mathbf{Z}^{1,new}$ are based on the same covariate values whose functional form may, however, be different in model (1) depending on the group pertinence.

In the *marginal approach*, $f_{g,new}$ equals to the marginal density of \mathbf{Y}^{new} , given the parameters of the g -th mixed model (1), i.e.,

$$f_{g,new} = f_g(\mathbf{y}^{new}) = \varphi(\mathbf{y}^{new} | \mathbf{X}^g \hat{\boldsymbol{\beta}}^g, \mathbf{Z}^g \hat{\mathbf{D}}^g \mathbf{Z}^{g'} + \hat{\sigma}^{g2} I_n). \quad (3)$$

In the *conditional approach*, $f_{g,new}$ equals to the conditional density of \mathbf{Y}^{new} , given the empirical Bayes estimates $\hat{\mathbf{b}}^g$ of random effects and parameters of the g -th mixed model (1), i.e.,

$$f_{g,new} = f_g(\mathbf{y}^{new} | \hat{\mathbf{b}}^g) = \varphi(\mathbf{y}^{new} | \mathbf{X}^g \hat{\boldsymbol{\beta}}^g + \mathbf{Z}^g \hat{\mathbf{b}}^g, \hat{\sigma}^{g2} I_n). \quad (4)$$

Finally, the *random effects approach* uses a density of random effects evaluated at the empirical Bayes estimate $\hat{\mathbf{b}}^g$ to compute posterior probabilities, i.e.,

$$f_{g,new} = f_g(\hat{\mathbf{b}}^g) = \varphi(\hat{\mathbf{b}}^g | \mathbf{0}, \hat{\mathbf{D}}^g). \quad (5)$$

By means of cross-validation, the three approaches are compared with respect to (1) sensitivity and specificity obtained with different cut-off values, (2) the success of classification performed either sequentially over time or at week 32 (the latest relevant moment for stopping the treatment since treatment is stopped anyway at week 52) and (3) the speed with which the posterior probabilities $\pi_{0,new}$ approach 1 in \mathcal{I}_0 group and to 0 in \mathcal{I}_1 group.

4 Results and Conclusions

In our model, a different mean evolution of \log_{10} (viral load) and \log (ALT) was allowed for different virus genotypes by inclusion of interaction terms in design matrices \mathbf{X}_i^g . Unsurprisingly, the quality of overall response prediction increases with the distinction of the mean profiles of \log_{10} (viral load)

and $\log(\text{ALT})$ between $R = 0$ and $R = 1$ group. For example, nicely separated mean profiles have been obtained for patients with genotype A virus, see Figure 1, where a crude classification rule based on a cut-off of 0.5 for the posterior probability applied only at week 32 led to 96.2%, 92.3% and 92.3% correctly classified patients from $R = 0$ group when marginal, conditional and random effect approach, respectively has been applied. Proportions of correctly classified patients from (a priori less prevalent) $R = 1$ group were 57.9%, 57.9%, 78.9%, respectively. On the other hand, much worse proportions of correctly classified patients (85.3%, 79.4%, 73.5%, respectively in $R = 0$ group and 33.3%, 33.3%, 41.7%, respectively) have been observed for patients with genotype D virus where the mean profiles

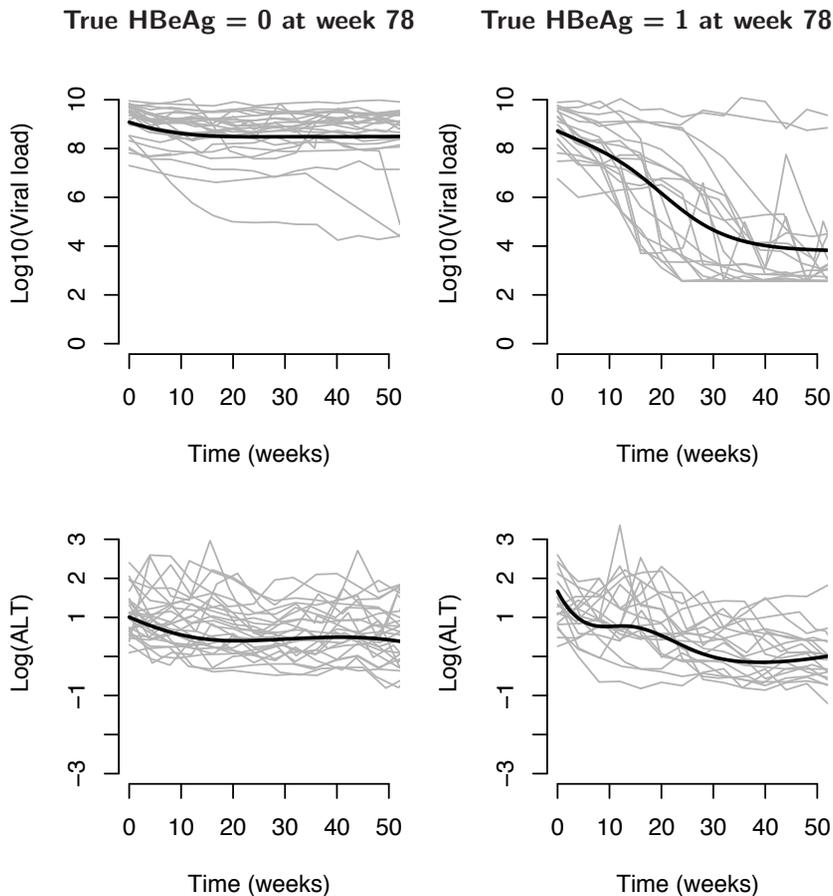


FIGURE 1. Observed profiles of $\log_{10}(\text{viral load})$ and $\log(\text{ALT})$ for patients with genotype A virus (grey lines) and fitted mean profiles for a median patient (black lines).

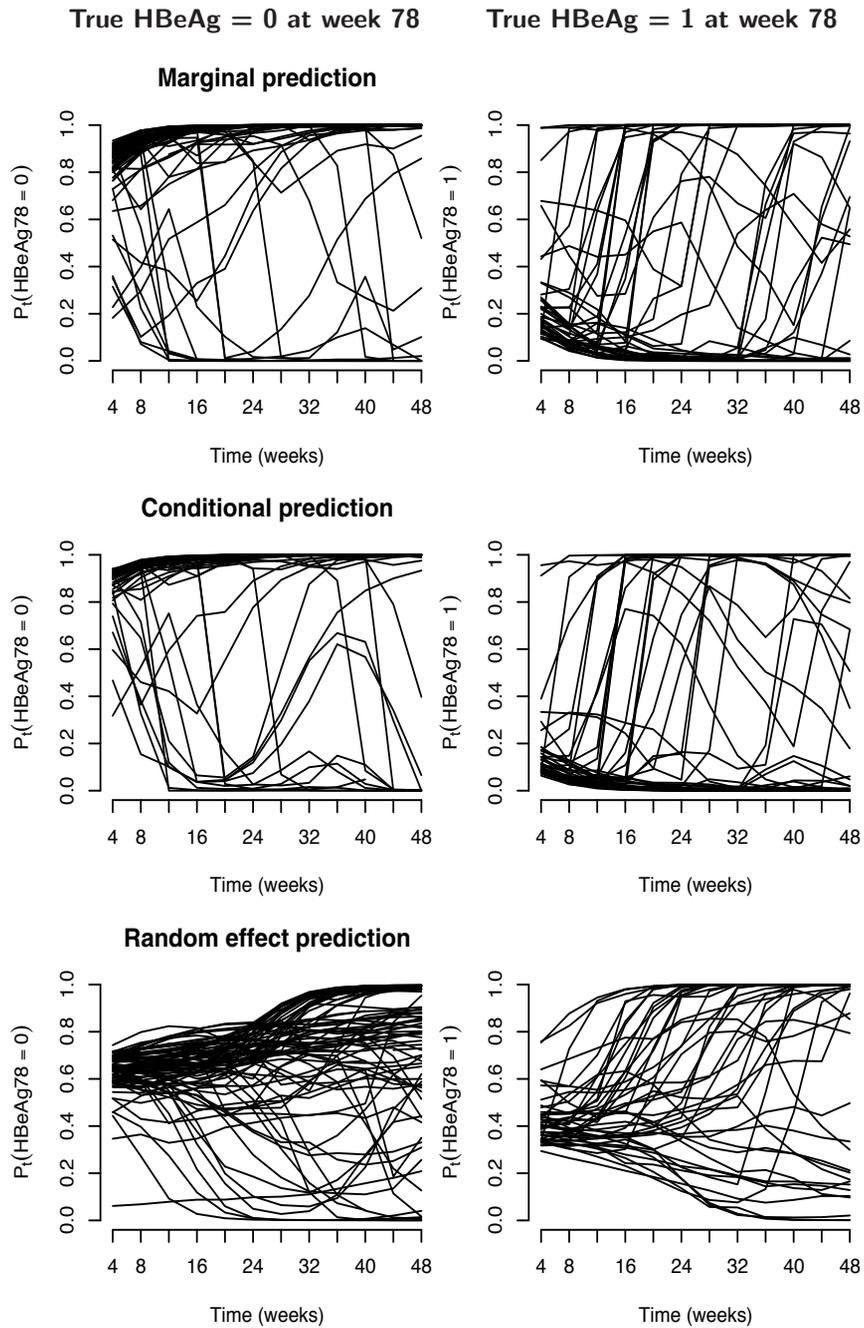


FIGURE 2. Evolution of posterior probabilities of belonging to correct HBeAg78 group over time. Left part: true HBeAg = 0 at week 78, right part: true HBeAg = 1 at week 78.

of longitudinal markers were not much separated.

The whole evolution of posterior probabilities over time is shown in Figure 2. It is seen that by week 20, marginal and conditional approaches would missclassify a high proportion of $HBeAg_{78} = 1$ patients. On the other hand, the random effects approach remains keeping posterior probabilities around the values of their prior values during the first few weeks and then let the posterior probabilities jump in the right direction in most cases. Hence it seems that the random effect approach is the most promising one. However, there is a space for improvement from methodological point of view, especially with respect to the assumed distribution of the random effects. We aim to investigate the models in which the normal distribution will be replaced by another more flexible distributional form. Further, our aim is to develop models in which we allow for discrete markers as well, i.e., models in which components of vector \mathbf{Y} can be mixed continuous and discrete. Possible improvements of this type will be discussed in more detail on the poster.

Acknowledgements

The first author acknowledge the support of 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

- Dierckx, P. (1993). *Curve and Surface Fitting with Splines*. Oxford: Clarendon.
- Janssen, H. L. A., van Zonneveld, M., Senturk, S., Akarca, U. S., Cakaloglu, Y., Simon, C., So, T. M. K., Gerken, G., de Man, R. A. and Niesters, H. G. M., Zondervan, P., Hansen, B., and Schalm, S. W. (2005). Pegylated interferon alfa-2b alone or in combination with lamivudine for HBeAg-positive chronic hepatitis B: a randomised trial. *Lancet*, **365**, 123–129.
- Morrell, C. H., Brant, L. J., and Sheng, S. (2007). Comparing approaches for predicting prostate cancer from longitudinal data. In *2007 Proceedings of the American Statistical Association*, Biometrics Section, pages 127–133. Alexandria: American Statistical Association.
- Morrell, C. H., Brant, L. J., Sheng, S., and Metter, E. J. (2005). Using multivariate mixed-effects models to predict prostate cancer In *2005 Proceedings of the American Statistical Association*, Biometrics Section, pages 332–337. Alexandria: American Statistical Association.

The iterative adjustment of the responses for the reduction of bias in binary regression models

Ioannis Kosmidis¹

¹ Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK

Abstract: The bias-reducing adjusted score functions (Firth, 1993, *Biometrika*) are studied for binomial-response generalized linear models. It is shown that the adjusted score functions imply a set of adjustments to the binomial responses and totals. An appropriate expression of the adjusted responses and totals is given, so that the adjusted data mimic the range of the binomial responses and totals. A procedure for obtaining the bias-reduced estimates is developed which relies on the iterative adjustment of the binomial responses and totals using ready maximum likelihood implementations. Furthermore, it is shown that the bias-reduced estimator, as the maximum likelihood estimator, is invariant to the representation of the binomial data. A complete enumeration study is used to demonstrate the superior statistical properties of the bias-reduced estimator to the maximum likelihood estimator.

Keywords: bias reduction, adjusted responses, adjusted score functions

1 Introduction

In statistical modelling, the additive adjustment of the binomial data by a constant a is a common practice, mainly for avoiding sparseness issues which may result to infinite maximum likelihood estimates and severe bias, or for, generally, improving the asymptotic behaviour of the estimators. Consider independent binomial random variables Y_1, \dots, Y_n with totals m_1, \dots, m_n and probabilities π_1, \dots, π_n and the logistic regression model,

$$\log\left(\frac{\pi_r}{1 - \pi_r}\right) = \eta_r = \sum_{t=1}^p \beta_t x_{rt} \quad (r = 1, \dots, n), \quad (1)$$

with x_{rt} the (r, t) th component of an $n \times p$ design matrix X and with β_1, \dots, β_p unknown parameters (an intercept can be included in the model by setting the components of a column of X to one).

Perhaps the most famous adjustment is the Haldane-Anscombe correction (Haldane, 1955; Anscombe, 1956), with $a = 1/2$, which results in an estimator of the log-odds $\log\{\pi/(1 - \pi)\}$ with bias of order $O(m^{-2})$ and has

the further advantage that the resultant estimate is finite. For the estimation of the parameters of a logistic regression model with $\eta_r = \beta_1 + \beta_2 x_r$ in (1), Hitchcock (1962) showed that the Haldane-Anscombe correction is not optimal in terms of the bias of the estimators and proposed $a = 1/4$ for $n = 3$ and no adjustment for $n > 3$. Hitchcock (1962) also noted that the first-order biases ($O(m^{-1})$ when $m = m_1 = \dots = m_n$) depend on the parameter values (see, Gart & Zweifel 1967, for a comparison of the above adjustments and some other adjustment schemes in terms of the bias of resultant log-odds estimators). In Gart et al. (1985), the Hitchcock (1962) proposal is verified for $n = 2, 3$ and 4 and refined for $n > 4$, demonstrating that for $n > 2$ there is not a universally optimal value of the constant a .

Clogg et al. (1991) presented a more sophisticated adjustment scheme for general logistic regressions where based on standard Bayesian arguments relating to the behaviour of the Jeffreys prior amongst every possible logistic regression, a was chosen to be $p \sum_{r=1}^n y_r / (n \sum_{r=1}^n m_r)$ and p/n was appended to the totals (see, also Rubin & Schenker 1987). The stated aim in Clogg et al. (1991) was not bias reduction but rather an applicable method of eliminating the possibility of infinite maximum likelihood estimates for the many logistic regressions which were involved in the large application that had been considered. Though the resultant estimator enjoys certain shrinkage properties which result in some reduction of the bias.

All the above adjustment schemes have the common property that a is constant with respect to the parameter vector $\beta = (\beta_1, \dots, \beta_p)$ and thus estimation can be conveniently performed by the following procedure:

1. adjust the responses by adding the constant a , and
2. proceed with usual estimation methods, treating the adjusted responses as actual.

However, because the adjustments are constants, the resultant estimators are generally not invariant under different representations of the data (for example, aggregated and disaggregated view), a desirable invariance property that the maximum likelihood estimator has. Furthermore, the first-order bias of the maximum likelihood estimator for logistic regressions generally depends on the parameter values (see Cordeiro and McCullagh, 1991, for explicit expressions) and thus there cannot be a universal constant a which always eliminates the first-order bias. Firth (1993), in his study of bias-reducing adjusted score functions, presented a parameter-dependent adjustment scheme for logistic regressions which eliminates the first-order bias of the estimator. That adjustment scheme consists of the addition of half a leverage and a leverage to the binomial responses and totals, respectively.

In the current note, the results in Firth (1993) for logistic regressions are extended in binomial-response generalized linear models with arbitrary link functions. The set of bias-reducing adjustment schemes is presented and an

appropriate expression of the adjusted responses and totals is given so that the adjusted data mimic the range of the binomial responses and totals ($0 \leq y_r \leq m_r$). An alternative to the modified iterative re-weighted least squares algorithm in Kosmidis & Firth (2008) is developed, where the solutions of the adjusted score equations can be obtained simply by the use of ready maximum likelihood implementations and appropriately adjusted responses and totals. Furthermore, it is shown that the resultant adjustment schemes result in estimates that are invariant to the representation of the binomial data. A complete enumeration study is used to demonstrate the finiteness and shrinkage properties of the bias-reduced estimator as well as its better performance in terms of bias and mean squared error over the maximum likelihood estimator. Furthermore, the effect of bias-reduction to the fitted probabilities is discussed.

2 Bias reducing adjustments to the score functions

Consider the same setup as for (1) but where the binomial probabilities are linked to the model parameters as

$$g(\pi_r) = \eta_r \quad (r = 1, \dots, n), \quad (2)$$

with $g(\cdot)$ a monotone function from $[0, 1]$ to the real line. According to the results in Kosmidis & Firth (2008, Section 4.1), a second-order unbiased estimator of β can be obtained by the solutions of the adjusted score equations $U_t^* = 0$ ($t = 1, \dots, p$), with

$$U_t^* = \sum_{r=1}^n \frac{w_r}{d_r} \left(y_r + \frac{1}{2} h_r \frac{d'_r}{w_r} - m_r \pi_r \right) x_{rt}, \quad (3)$$

where $d_r = m_r d\pi_r/d\eta_r$, $d'_r = m_r d^2\pi_r/d\eta_r^2$, $w_r = d_r^2/\{m_r\pi_r(1-\pi_r)\}$ is the r th quadratic weight and h_r is the r th diagonal element of the hat matrix $H = X(X^T W X)^{-1} X^T W$, with W the diagonal matrix with non-zero elements w_r ($r = 1, \dots, n$). The above adjusted score functions suggest appending $h_r d'_r/(2w_r)$ to the binomial response y_r ($r = 1, \dots, n$). Kosmidis & Firth (2008) give the form of the adjusted responses for binomial-response generalized linear models for some well-known link functions.

3 Adjustment of the binomial responses and totals

3.1 An appropriate pseudo-data representation

A convenient way of solving the adjusted-score equations would be to use the adjusted responses in ready maximum likelihood implementations, iteratively. Nevertheless, a practical issue that can arise relates to the sign

of $h_r d'_r / w_r$ (or simply the sign of d'_r) which can result in negative adjusted responses or adjusted responses greater than the binomial totals, violating the range of the actual data ($0 \leq y_r \leq m_r$). Fortunately, this issue can be resolved with simple algebraic manipulation.

Dropping the subject index r , a pseudo-data representation is defined as the pair $\{y^*, m^*\}$, with y^* the adjusted response and m^* the adjusted binomial total. By this definition, the apparent pseudo-data representation suggested by (3) is $\{y + hd' / (2w), m\}$. Nevertheless, the form of (3) suggests that there is a countable set of equivalent pseudo-data representations, equivalent in the sense that if the actual responses and totals are replaced with $\{y^*, m^*\}$ in the likelihood equations then the adjusted score equations result. Any pseudo-data representations in this set can be resulted from any other by the operations of adding and subtracting a quantity to either the adjusted responses or the adjusted totals, and of moving summands from the adjusted responses to the adjusted totals after division with $-\pi$.

Within this set of pseudo-data representations consider the ones which have the form

$$\left\{ y + h \frac{d'}{2w} + h\pi b, \quad m + hb \right\}, \tag{4}$$

with b some function of η . Substituting for w , because $0 \leq h \leq 1$ and $0 \leq y \leq m$, a sufficient condition for the adjusted responses and totals to mimic the range of the actual responses and totals (ie. $0 \leq y^* \leq m^*$ ($r = 1, \dots, n$)) is that $b \geq md'(\pi - 1)/(2d^2)$ and $b \geq md'\pi/(2d^2)$ are satisfied simultaneously. Thus, because $0 \leq \pi \leq 1$, the requirement $0 \leq y^* \leq m^*$ can be met, for example, if $b = md'(\pi - 1)/(2d^2) + 1/2$ for $d' \leq 0$ and if $b = md'\pi/(2d^2) + 1/2$ for $d' > 0$. Hence, b can be set to $md'(\pi - I_{d' \leq 0})/(2d^2) + 1/2$. Substituting in (4), we obtain the pseudo-data representation

$$\left\{ y + \frac{1}{2}h\pi \left(1 + \frac{md'}{d^2} I_{d' > 0} \right), \quad m + \frac{1}{2}h \left(1 + \frac{md'}{d^2} (\pi - I_{d' \leq 0}) \right) \right\}, \tag{5}$$

where I_E is 1 if E holds and 0 otherwise.

3.2 Local maximum likelihood fits on pseudo-data representations

In light of 5, the bias-reduced estimates can be obtained by an iterative adjustment procedure where the $(j + 1)th$ iteration is as follows:

- i) Update to $\{y_{r,(j+1)}^*, m_{r,(j+1)}^*\}$ according to (5) evaluating all the quantities involved at the estimates $\beta_{(j)}$ from the jth iteration.
- ii) Use maximum likelihood to fit model (2) with responses $y_{r,(j+1)}^*$ and totals $m_{r,(j+1)}^*$ ($r = 1, \dots, n$), using $\beta_{(j)}$ as starting value.

If the maximum likelihood estimates are finite, they provide sufficiently good starting values for the above iteration. Otherwise, the iteration can start at the maximum likelihood estimates obtained after the addition of a constant $a > 0$ to the responses and $2a$ to the totals.

Furthermore, the condition $\sum_{t=1}^p |U_t^*(\beta_{(j+1)})| \leq \epsilon$, $\epsilon > 0$ can be used as a general convergence criterion of the procedure.

An implementation of the above procedure can be found in the *brglm* R package (Kosmidis, 2007a).

3.3 Invariance of the estimates under the structure of the data

It is always possible to represent Y_r as $\sum_{s=1}^{k_r} Z_{rs}$, where Z_{r1}, \dots, Z_{rk_r} are independent binomial random variables each with probability of success π_r and totals l_{r1}, \dots, l_{rk_r} , respectively, with $m_r = \sum_{s=1}^{k_r} l_{rs}$ ($r = 1, \dots, n$). By this construction, in the presence of a covariate vector x_r for each observation y_r , the data for a binomial-response generalized linear model can be represented in two equivalent ways, (y_r, m_r, x_r) ($r = 1, \dots, n$) and (z_{rs}, l_{rs}, x_r) ($r = 1, \dots, n; s = 1, \dots, k_r$). The maximum likelihood estimator of the model parameters is invariant to the choice of either representation and, in contrast to constant adjustment schemes, the bias-reduced estimator also has the same invariance property.

To show that, denote z_{rs}^* and l_{rs}^* ($r = 1, \dots, n; s = 1, \dots, k_r$), the adjusted responses and totals, respectively. Note that the bias-reducing pseudo-data representations have the generic form $\{z_{rs} + \tilde{h}_{rs}q_r, m_{rs} + \tilde{h}_{rs}v_r\}$ where $q_r \equiv q(\pi_r)$ and $v_r \equiv v(\pi_r)$ and \tilde{h}_{rs} is the generic diagonal element of the hat matrix. Note that, q_r and v_r depend solely on π_r and, also, a simple calculation can show that $\sum_{s=1}^{k_r} \tilde{h}_{rs} = h_r$ ($r = 1, \dots, n$). Hence, $\sum_{s=1}^{k_r} z_{rs}^* = y_r^*$ and $\sum_{s=1}^{k_r} l_{rs}^* = m_r^*$ ($r = 1, \dots, n$) and because the adjusted score functions result by the replacement of the actual responses and totals by their adjusted versions, the bias-reduced estimates are invariant to the structure of the binomial data.

4 Demonstration of the properties of the bias-reduced estimator

For the demonstration of the properties of the bias-reduced estimator let $n = 5$ and $m_r = m$ ($r = 1, \dots, 5$). Furthermore, consider the model (2) with $\eta_r = \beta_1 + \beta_2 x_r$ and let $x = (-2, -1, 0, 1, 2)$. For $m = 4, 8, 16$ and for the logit, probit and complementary log-log link functions, the bias and mean squared error of the bias-reduced estimator, as well as, the coverage of the nominally 95% Wald-type confidence interval were calculated through complete enumeration when the true parameter values are $\beta_1 = -1$ and $\beta_2 = 1.5$. This calculation is possible, because for every data set and every link function, the bias-reduced estimates were finite. Nevertheless, the

TABLE 1. The results of the complete enumeration. The parenthesized probabilities refer to the event of encountering at least one infinite ML estimate for the corresponding m and link function. All the quantities for the ML estimator are calculated conditionally on the finiteness of its components.

Link	m	Parameter	Bias ($\times 10^2$)		MSE ($\times 10$)		Coverage	
			ML	BR	ML	BR	ML	BR
logit	4	β_1	-8.79	0.52	5.84	6.07	0.971	0.972
	(0.1621)	β_2	14.44	-0.13	3.62	4.73	0.960	0.939
	8	β_1	-12.94	-0.68	3.93	3.11	0.972	0.964
	(0.0171)	β_2	20.17	1.11	3.70	2.68	0.972	0.942
	16	β_1	-7.00	-0.19	1.75	1.42	0.961	0.957
	(0.0002)	β_2	10.55	0.29	1.59	1.16	0.960	0.948
probit	4	β_1	17.89	13.54	1.44	2.61	0.968	0.911
	(0.5475)	β_2	-18.84	-16.93	0.98	3.07	0.960	0.897
	8	β_1	0.80	3.24	1.07	1.82	0.964	0.938
	(0.2296)	β_2	6.08	-3.81	1.26	2.13	0.972	0.908
	16	β_1	-7.06	0.24	1.03	1.08	0.974	0.949
	(0.0411)	β_2	12.54	-0.17	1.39	1.22	0.973	0.933
cloglog	4	β_1	2.97	3.18	2.97	3.07	0.959	0.962
	(0.3732)	β_2	-2.93	-12.97	1.35	3.51	0.955	0.880
	8	β_1	-8.42	0.84	2.49	1.89	0.962	0.953
	(0.1000)	β_2	15.63	-5.40	2.33	2.36	0.972	0.906
	16	β_1	-6.45	0.17	1.32	0.98	0.964	0.957
	(0.0071)	β_2	13.13	-1.74	1.60	1.23	0.965	0.921

ML: maximum likelihood; BR: bias reduction.

maximum likelihood estimator had at least one infinite components with positive probability. Thus the corresponding quantities for the maximum likelihood estimator are undefined and were only calculated conditionally on the finiteness of both its components.

The results are shown in Table 1. Direct comparison of the conditional moments and coverage for the maximum likelihood estimator with the corresponding unconditional quantities for the bias-reduced estimator is misleading, in the direction of favouring the method of maximum likelihood. Despite this reservation, according to the results in Table 1, in cases where the probability of infinite maximum likelihood estimates is small or moderate, the bias-reduced estimator has bias and mean squared error properties that are better, even, than the corresponding conditional quantities for maximum likelihood.

In Figure 1, using the results of the complete enumeration for the complementary log log link with $m = 4$, the fitted probabilities based on the bias-reduced estimates are plotted against the fitted probabilities based on the maximum likelihood estimates. The shrinkage of the former towards

$1 - \exp(-1) \approx 0.632$ is apparent. Correspondingly, the bias-reduced estimates shrink towards the origin of the scale of the linear predictor relative to the maximum likelihood estimates. This behaviour is typical for binomial-response generalized linear models; for the probit and logit model the fitted probabilities shrink towards 0.5 (see, also, Kosmidis, 2007b). From the results in Table 1, the coverage of the nominally 95% Wald-type confidence intervals for the bias-reduced estimator is poor in this setting. A reason for this is that the performance of the intervals is studied for extreme true parameter values; the finiteness and shrinkage properties of the bias-reduced estimator result in smaller estimated asymptotic standard errors (square roots of the diagonal of the Fisher information) so that the resultant Wald-type confidence intervals are short in length and do not cover extreme effects with sufficiently high probability. In contrast, for $m = 4$ and for true parameter values $\beta_1 = -1$ and $\beta_2 = 0.5$, the Wald-type confidence interval for the bias-reduced estimator of β_2 has better coverage behaviour with coverage 0.969 for the probit link and 0.971 for the complementary log log link. Typically, as the sample size increases the coverage tends to the nominal level.

5 Discussion

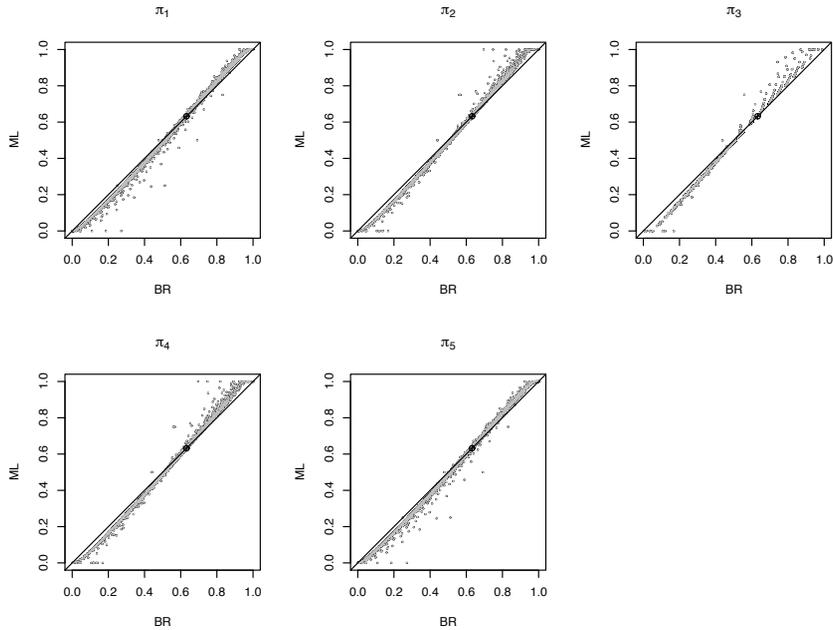
For binomial-response generalized linear models, it has been shown how the bias reduction method in Firth (1993) can be implemented by iteratively adjusting the binomial responses and totals and using ready maximum likelihood implementations.

Furthermore, it has been shown that, as the maximum likelihood estimator, the bias-reduced estimator is invariant to the representation of the binomial data. In addition, contrastingly to the maximum likelihood estimates, as has been demonstrated through complete enumeration, the bias-reduced estimates are always finite and because of their improved statistical properties, their routine use in applications is appealing.

Nevertheless, Wald-type approximate confidence intervals for the bias-reduced estimator can have bad coverage properties. In the case of logistic regression, the adjusted score functions correspond to the penalization of the likelihood by Jeffreys invariant prior (Firth, 1993). Heinze & Schemper (2002) used this fact and illustrated that approximate confidence intervals based on the profiles of the penalized likelihood can have better coverage properties than Wald-type approximate confidence intervals. However, according to Kosmidis & Firth (2008, Theorem 1), the adjusted score functions for binomial-response generalized linear models with non-logistic link, do not generally admit a penalized likelihood interpretation. In that cases the adjusted-score statistic

$$U_t^*(\hat{\beta}_1, \dots, \hat{\beta}_{t-1}, \beta_t, \hat{\beta}_{t+1}, \dots, \hat{\beta}_p)^2 F^{tt}(\hat{\beta}_1, \dots, \hat{\beta}_{t-1}, \beta_t, \hat{\beta}_{t+1}, \dots, \hat{\beta}_p)$$

FIGURE 1. Fitted probabilities based on the bias-reduced estimates against the fitted probabilities based on the maximum likelihood estimates for the cloglog link with $m = 4$. The marked point on the plots is (c, c) , where $c = 1 - \exp(-1)$.



could be used for the construction of confidence intervals for the parameter β_t ($t = 1, \dots, p$). Here, $\hat{\beta}_u$ ($u = 1, \dots, t-1, t+1, \dots, p$) are the bias-reduced estimates when the t th component of the parameter vector is fixed at β_t and F^{tt} is the (t, t) th component of the inverse Fisher information. Because $U_t^* = U_t + A_t$, where U_t is the t th component of the score vector and A_t is $O(1)$ as the sample size increases, the adjusted-score statistic is asymptotically distributed according to a chi-squared distribution with one degree of freedom.

References

- Agresti, A. (2002). *Categorical Data Analysis*. New York: Wiley.
- Anscombe (1956). On estimating binomial response relations. *Biometrika*, **43**, 461–464.
- Clogg, C. C., D. B. Rubin, N. Schenker, B. Schultz, and L. Weidman (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association*, **86**, 68–78.

- Cordeiro, G. M. and P. McCullagh (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society, Series B: Methodological*, **53**, 629–643.
- Firth, D. (1992a). Bias reduction, the Jeffreys prior and GLIM. In L. Fahrmeir, B. Francis, R. Gilchrist, and G. Tutz (Eds.), *Advances in GLIM and Statistical Modelling: Proceedings of the GLIM 92 Conference, Munich*, New York, pp. 91–100. Springer.
- Firth, D. (1992b). Generalized linear models and Jeffreys priors: An iterative generalized least-squares approach. In Y. Dodge and J. Whittaker (Eds.), *Computational Statistics I*, Heidelberg. Physica-Verlag.
- Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27–38.
- Gart, J. J., H. M. Pettigrew, and D. G. Thomas (1985). The effect of bias, variance estimation, skewness and kurtosis of the empirical logit on weighted least squares analyses. *Biometrika*, **72**, 179–190.
- Gart, J. J. and J. R. Zweifel (1967). On the bias of various estimators of the logit and its variance with application to quantal bioassay. *Biometrika*, **54**, 181–187.
- Haldane, J. (1955). The estimation of the logarithm of a ratio of frequencies. *Annals of Human Genetics*, **20**, 309–311.
- Heinze, G. and M. Schemper (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, **21**, 2409–2419.
- Hitchcock, S. E. (1962). A note on the estimation of parameters of the logistic function using the minimum logit χ^2 method. *Biometrika*, **49**, 250–252.
- Kosmidis, I. (2007a). brglm: Bias reduction in binomial-response GLMs. R package. <http://go.warwick.ac.uk/kosmidis/software>.
- Kosmidis, I. (2007b). *Bias reduction in exponential family nonlinear models*. Ph. D. thesis, Department of Statistics, University of Warwick.
- Kosmidis, I. and D. Firth (2008). Bias reduction in exponential family non-linear models. Technical Report 8-5, CRiSM working paper series, University of Warwick.
- Rubin, D. B. and N. Schenker (1987). Logit-based interval estimation for binomial data using the Jeffreys prior. *Sociological Methodology*, **17**, 131–144.

Nested B -spline bases: an efficient method for spatio-temporal smoothing

Dae-Jin Lee¹ and María Durbán¹

¹ Department of Statistics, Universidad Carlos III de Madrid, SPAIN.
e-mail: dae-jin.lee@uc3m.es and mdurban@est-econ.uc3m.es

Abstract: The methodology developed for the analysis of spatio-temporal data is often constrained by the size of the data sets, and most models impose unrealistic constraints on the data in order to be fitted on a reasonable amount of time. In the context of environmental studies, when data present a strong seasonal trend, the size of the basis needed to capture the temporal trend is large and, as a consequence, the estimation of the spatio-temporal interaction is computationally inefficient. We propose the use of the mixed model representation of P -splines to fit ANOVA-type spatio-temporal models, and reduce the computational burden by using smaller bases for the interaction term. These basis functions are easily calculated by using an appropriate number of knots, they preserve the hierarchical nature of the models, and dramatically reduce the number of parameters fitted. We illustrate the advantage of this method with the analysis of average monthly temperature data from 136 U.S. cities.

Keywords: spatio-temporal data; P -splines; nested bases.

1 P -splines for spatio-temporal smoothing

Most of the common approaches in spatio-temporal smoothing assume an additive model with two components: a two-dimensional term for the spatial surface and a one-dimensional term for the temporal dimension (Kneib 2006; MacNab, 2007). Therefore, they impose a separable structure for the spatio-temporal covariance that, in many cases, will not represent the real structure of the data. As an alternative, Lee and Durbán (2008) proposed a class of non-separable models of the form

$$\mathbf{y} = f(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_t) + \epsilon = \mathbf{B}\theta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}), \quad (1)$$

where the data are located in n geographical locations, $s = (\mathbf{x}_1, \mathbf{x}_2)$, and measured over t time periods \mathbf{x}_t . The regression basis for a $3d$ interaction model is

$$\mathbf{B} = (\mathbf{B}_1 \square \mathbf{B}_2)_s \otimes \mathbf{B}_t = \mathbf{B}_s \otimes \mathbf{B}_t, \quad nt \times c_1 c_2 c_3, \quad (2)$$

where \mathbf{B}_1 , \mathbf{B}_2 and \mathbf{B}_t are the marginal B -spline basis of dimensions $n \times c_1$, $n \times c_2$ and $t \times c_3$ respectively, θ is the vector of regression parameters, and symbol \square represents the row-tensor Kronecker product.

Smoothness is imposed via the penalty matrix \mathbf{P} based on second order difference matrices \mathbf{D}_1 , \mathbf{D}_2 and \mathbf{D}_t . The penalty term in 3-dimensions is:

$$\mathbf{P} = \lambda_1 \mathbf{D}'_1 \mathbf{D}_1 \otimes \mathbf{I}_{c_2} \otimes \mathbf{I}_{c_3} + \lambda_2 \mathbf{I}_{c_1} \otimes \mathbf{D}'_2 \mathbf{D}_2 \otimes \mathbf{I}_{c_3} + \lambda_t \mathbf{I}_{c_1} \otimes \mathbf{I}_{c_2} \otimes \mathbf{D}'_t \mathbf{D}_t, \quad (3)$$

this penalty allows spatial *anisotropy* by considering a different amount of smoothing for longitude and latitude ($\lambda_1 \neq \lambda_2$) and for the temporal component (λ_t).

However, in most situations, we are interested on identifying the spatial and temporal component, as well as the interaction, and therefore, a more adequate model would be:

$$\mathbf{y} = f_s(\mathbf{x}_1, \mathbf{x}_2) + f_t(\mathbf{x}_t) + f_{s,t}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_t) + \epsilon, \quad (4)$$

with basis functions defined by blocks as:

$$\mathbf{B} = [\mathbf{B}_s \otimes \mathbf{1}_t : \mathbf{1}_n \otimes \mathbf{B}_t : \mathbf{B}_s \otimes \mathbf{B}_t], \quad (5)$$

with $\mathbf{1}_n$ and $\mathbf{1}_t$ are column vectors of ones' of length n and t respectively, where each block of (5) corresponds to each of the smooth functions defined in (4). And block-diagonal penalty:

$$\mathbf{P}^* = \begin{bmatrix} \tau_1 \mathbf{D}'_1 \mathbf{D}_1 \otimes \mathbf{I}_{c_2} + \tau_2 \mathbf{I}_{c_1} \otimes \mathbf{D}'_2 \mathbf{D}_2 & & \\ & \tau_t \mathbf{D}'_t \mathbf{D}_t & \\ & & \mathbf{P} \end{bmatrix}, \quad (6)$$

with \mathbf{P} defined in (3).

Model (4) could be seen as a particular case of the ANOVA-type models proposed by Chen (1993), but with the advantage of using low-rank smoothers. It is well known that these models have problems with the identifiability of the different terms, and it is necessary to impose constraints. This is due to the fact that the basis for the interaction term, $\mathbf{B}_s \otimes \mathbf{B}_t$, span the space of functions fitted by the individual basis, \mathbf{B}_s and \mathbf{B}_t , and therefore matrix (5) is not of full column rank ($\text{rank}(\mathbf{B}) = c_1 c_2 c_3$).

Lee and Durbán (2008) used a reparameterization of the model (based on the mixed models representation of P -splines) that yielded a very simple method for identifying the columns that are linearly dependent. Furthermore, they showed that removing the linearly dependent columns is equivalent to fitting constraints on the parameters which are equivalent to those applied in a factorial design. This method of imposing the identifiability constraints constructs models that are nested and can be easily compared.

These models are also fitted taking advantage of the array structure of the space-time interaction and using the GLAM algorithms proposed by Currie et al. (2006). However, when the number of time points is large, the number of parameters fitted could increase rapidly. A possible solution is the use of nested B -splines bases in the construction of (5).

2 Nested B -spline bases for computational efficiency

Environmental data often present a strong seasonal trend, and the size of the basis \mathbf{B}_t in (5) has to be large (between 20 and 40 equidistant knots) in order to have enough degrees of freedom to capture the temporal structure. As a consequence, the number of parameters in the interaction $\mathbf{B}_s \otimes \mathbf{B}_t$ could easily be of the order of thousands, and the computational burden prohibitive. A first solution would be just to use a smaller basis for the interaction term: take $\tilde{\mathbf{B}}_t$ such that $\text{rank}(\tilde{\mathbf{B}}_t) < \text{rank}(\mathbf{B}_t)$. This has two advantages: it reduces the total number of parameters, and will estimate a smoother interaction trend, avoiding the possibility of capturing the short-term correlation that might also be present in the data.

However, taking a reduced basis of arbitrary size will yield a model that will not be nested on the additive model $f_s(\mathbf{x}_1, \mathbf{x}_2) + f_t(\mathbf{x}_t)$, and so, the comparison between an additive and an interaction model will not be straightforward. The solution is to use nested bases for the interaction term, i.e., basis such that the space spanned by $\tilde{\mathbf{B}}_t$, is a subset of the space spanned by \mathbf{B}_t , and so, the hierarchical nature of the models is preserved, the identifiability constraints remain the same, and the number of parameters is

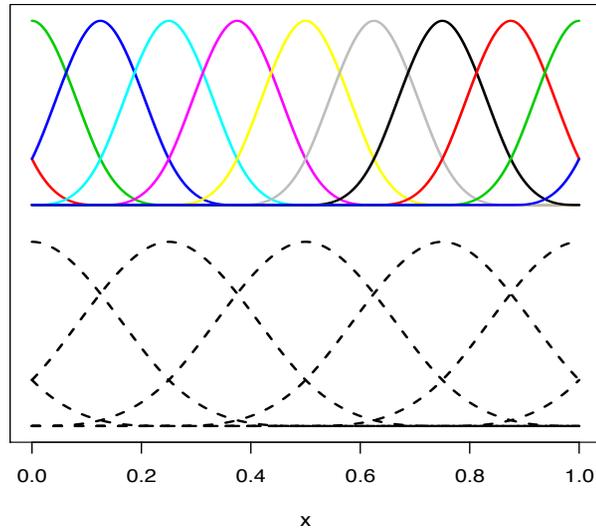


FIGURE 1. Nested B -spline bases with 8 knots (top) and 4 knots (bottom).

greatly reduced. The way to ensure that the new basis is nested on the original is to use a number of knots that is a divisor of the number of knots used in the original basis:

$$\#\text{knots}(\tilde{\mathbf{B}}_t) = \frac{\#\text{knots}(\mathbf{B}_t)}{d} \Rightarrow \text{span}(\tilde{\mathbf{B}}_t) \subset \text{span}(\mathbf{B}_t),$$

and d is any divisor of the number of knots used to construct \mathbf{B}_t (Figure (1) shows an example of basis with 8 and 4 knots). Therefore, a reduced basis to fit model (4) is:

$$\mathbf{B} = [\mathbf{B}_s \otimes \mathbf{1}_t : \mathbf{1}_n \otimes \mathbf{B}_t : \mathbf{B}_s \otimes \tilde{\mathbf{B}}_t]. \quad (7)$$

3 Application to U.S. temperature data

We apply the methodology proposed to the analysis of monthly average temperatures in $^{\circ}\text{F}$ for 136 cities across the U.S. between January 1995 and December 2004. The total number of observations was 16320. The R package `mgcv` developed by Wood (2009) fit similar type of models by means of the function `gamm`. However, due to storage and memory limitations involving the use of Kronecker products, it results unfeasible to obtain a satisfactory solution with a flexible enough size of marginal bases. In contrast, we use GLAM methods which allow us to store the data and model matrices more efficiently and speed up the calculations.

We considered 10 equidistant knots for each \mathbf{x}_1 and \mathbf{x}_2 coordinates, to cover the spatial domain. For the time trend, we chose 4 knots per year, with a total of 30 knots across \mathbf{x}_t (otherwise the seasonal effect would not be captured). Fitting the model (4) with a non-nested basis (5) led to a total of 5779 parameters, and the size of the matrices involved made the fit of the model computationally very intensive in standard requirements PCs. Decreasing the number of knots in the temporal dimension lead us to over-smoothing the seasonal trend. Figure 2, shows the smoothed time trends estimated using model (4) with different number of knots in the construction of \mathbf{B}_t and illustrates the need of choosing a large marginal basis for time. In Figure 3 the smooth spatial trend is shown.

For the nested bases approach (7), we construct $\tilde{\mathbf{B}}_t$, with 15, 10 and 6 knots. Table 1, shows the % of reduction achieved by using the nested basis in terms of total number of parameters estimated and cpu time required to perform the estimation algorithms.

We compared the performance of the models in terms of the Akaike Information Criteria (AIC), and the model degrees of freedom, measured as $\text{df} = \text{trace}(\mathbf{H})$, where \mathbf{H} is the “*hat matrix*”. The results obtained are summarized in Table 2, although the AIC increases with the reduction of the number of parameters, this reduction supposed a decrease of a 8% and did not significantly affect the goodness of fit of the model. Therefore, the selection of a more parsimonious nested model was a reasonable choice.

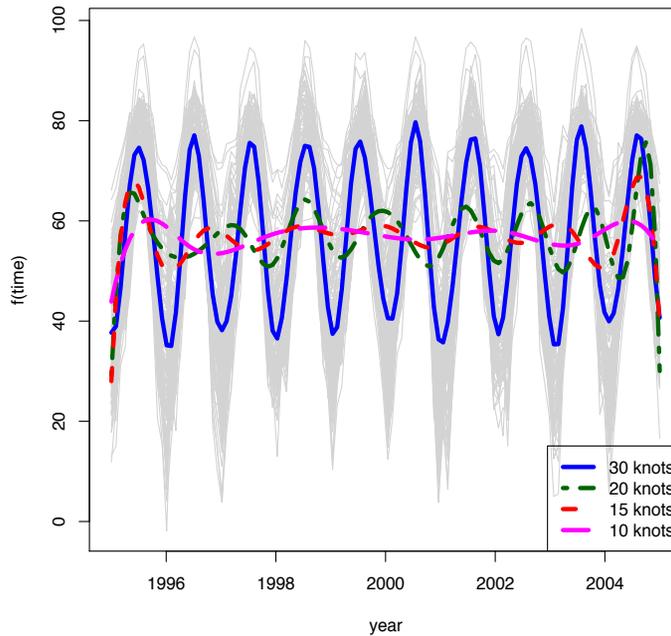


FIGURE 2. Monthly average temperature data ($^{\circ}\text{F}$) in 136 U.S. monitoring stations. Figure shows the smooth time trend $f_t(\mathbf{x}_t)$ with different number of knots (30,20,15, and 10). Less than 30 knots does not capture the seasonal effect.

TABLE 1. Summary of nested models with reduced number of knots, respect to non-nested model with 30 knots in \mathbf{B}_t and 5779 parameters.

# param.	# reduced knots	reduction (%)	cpu time reduction (%)
3075	15	47%	82%
2399	10	58%	94%
1723	6	70%	97%

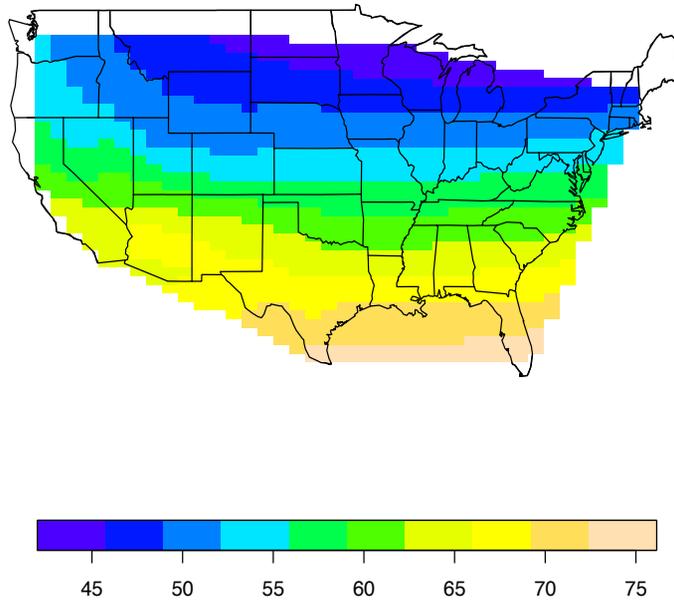
4 Concluding remarks

We have introduced a new computationally efficient method of spatio-temporal data smoothing. We proposed a way to reduce the computation time by the use of lower dimensional (nested) basis for the space-time interaction. The spatial and temporal components are captured by the remaining smooth terms in the ANOVA formulation of the model. The use of the nested basis has no significant loss in model performance, but a great gain in terms of computational efficiency.

This approach can also be extended to consider nested bases for the spatial

TABLE 2. Comparison of AIC and df of non-nested and nested basis models.

Model basis	AIC	df
non-nested	45424.37	878.18
nested		
15 knots	49374.31	158.35
10 knots	49565.24	147.84
6 knots	49648.49	109.40

FIGURE 3. Smooth spatial surface $f_s(\mathbf{x}_1, \mathbf{x}_2)$.

component, and define the space-time interaction basis as $\tilde{\mathbf{B}}_s \otimes \tilde{\mathbf{B}}_t$, where $\tilde{\mathbf{B}}_s$ is constructed from a reduced set of knots of the marginal basis of \mathbf{x}_1 and \mathbf{x}_2 . Depending on the spatial data structure, it may result necessary to increase the number of knots in the spatial smooth term, and use a lower dimension basis for the interaction and avoid computational complexity.

Acknowledgments: The research of Dae-Jin Lee and María Durbán was supported by Spanish Ministry of Education and Science under project MTM2008-02901.

References

- Chen, Z. (1993). Fitting Multivariate Regression Functions by Interactions Spline Models. *J. R. Statist. Soc. B*, **55**, 473-491.
- 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.
- Kneib, T. and Fahrmeir, L. (2006). Structured Additive Regression for Categorical Space-Time Data: A Mixed Model Approach. *Biometrics*, **62**, 109-118.
- Lee, D.-J. and Durbán, M. (2008). P-spline ANOVA-Type Interaction Models for Spatio-Temporal Smoothing *Proceedings of the 23rd International workshop on Statistical Modelling*, 315-320.
- MacNab, Y. C. (2007). Spline Smoothing in Bayesian Disease Mapping. *Environmetrics*, **18**, 727-744.
- Wood, S. N. (2006). Low-Rank Scale-Invariant Tensor Product Smooths for Generalized Additive Mixed Models. *Biometrics*, **62**, 1025-1036.
- R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Wood, S. N. (2009). *mgcv*: GAMs with GCV smoothness estimation and GAMMs by REML/PQL. In *R Package Version 1.5.2*.

Nonparametric Detection of Outliers in Multivariate Data Streams

Dongyu Lin¹, Shankar Krishnan² and Tamraparni Dasu²

¹ Department of Statistics, the Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA

² AT&T Labs - Research, 180 Park Avenue, Florham Park, NJ 07932, USA

Abstract: Data streams are dynamic, with frequent distributional changes. Detecting outliers in such data sources can often be challenging. A desired property of streaming outlier detection is to adapt naturally to such changes. In this paper, we propose a nonparametric method for detecting and understanding outliers in multivariate data streams. Our algorithm is a distance-based approach; it maintains an empirical distribution of signed distances and computes quantiles in a streaming fashion. In addition, we define an outlier curve that tracks the accumulation of observed outliers. The changes in this curve provide quantitative insights into the distributional changes in the stream. We have applied our technique to both synthetic and real-world multivariate data successfully.

Keywords: change detection; data streams; signed distance; outlier curve

1 Introduction

Change detection in data streams has important scientific, industrial and financial applications, for example in sensor networks, server usage monitoring, ATM transactions and other domains. Outlier detection is a well understood problem. However, the transient nature of data streams poses special challenges to change detection. Data streams are constantly changing, with high rates of accumulation and unknown distributional properties. An algorithm can access only a small slice of data at any given time, with no option of storing data for future reference [Gaber *et al.*, Babcock *et al.*]. These constraints require that change detection algorithms function within acceptable limits of storage and computational complexity, and reflect and adapt to the dynamic changes in a computationally efficient manner.

Classic outlier detection methods date back to control charts. Most known methods are univariate, ranging from parametric, model-based solutions to computational, data driven approaches. There are, however, several concerns while applying univariate methods to multivariate data. First of all, univariate tests do not capture inter-relationships between the variables and might result in incorrect identification of outliers. Outcome of univariate outlier detection methods need to be weighted and combined into a single

criterion. In addition, the problem of “multiple testing” that arises when the same data set is used to test multiple hypotheses in individual variables, might reduce the statistical power of the overall tests [Benjaminini *et al.*].

In statistical literature, multivariate outlier detection methods can be broadly classified as distance-based methods or projection pursuit methods. Both classes of methods are mostly designed for static data sets and are not easily extended to a streaming context.

Distance-based methods aim to detect outliers by computing a measure of how far a particular point is from the center of the data. This outlyingness is often based on a robust version of the Mahalanobis distance. Robust notions of the data center include coordinate-wise and L_1 median [Hossjer *et al.*]; distance-based algorithms that pursue robust estimation of the covariance matrix include the OGK estimate [Maronna *et al.*], the minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) [Rousseeuw *et al.*].

In contrast to distance-based procedures are projection pursuit methods [Donoho, Huber, Pena *et al.*, Galeano *et al.*]. These methods are characterized by finding suitable projections of the data in which the outliers are readily apparent. Free of distributional assumptions, projection pursuit techniques are applicable in diverse data situations [Filzmoser *et al.*], although at the expense of high computational cost.

Some nonparametric methods rely on the ranking of the data. Data depth is often used for this purpose in high dimensions. Outliers have low values of data depth while deeply embedded points like the median have higher data depth. Examples of nonparametric outlier detection methods include data depth [Liu *et al.*], regression depth [Rousseeuw *et al.*], and depth contours [Miller *et al.*]. However, most definitions of depth are computationally infeasible.

Given the dynamic nature of data streams, we believe that an outlier detection method must (a) adapt to the changing distribution and (b) distinguish between outliers that are a part of the changing distribution and those that are one-shot aberrations. In this paper, we propose a method that automatically adapts to changes, captures global and local shifts, and characterizes the nature of the changes in the data stream.

Briefly, we estimate the expected behavior of the data stream at time t using multivariate summaries and use a signed distance to capture deviation from expected behavior. The signed distance provides greater granularity for detecting outliers, particularly in skewed distributions. Quantiles of the signed distance are used as thresholds for outlier detection. A novel but critical component of our technique is an outlier curve OC_t that tracks the cumulative count of the outliers. Our method works on a wide range of

multivariate data streams, including *i.i.d.* sequences and time series. There are no distributional or model assumptions, hence it is widely applicable. It is also transparent, and easy to interpret and use.

2 Our Approach

Generally speaking, our approach is a distance-based streaming method. We leverage the adaptive nature of moving average techniques to compute the expected behavior of the data streams. The multivariate exponentially weighted moving average (EWMA) of $\mathbf{x}_t \in \mathbb{R}^d$ is defined recursively as

$$\boldsymbol{\mu}_t = \lambda \cdot \mathbf{x}_t + (1 - \lambda) \cdot \boldsymbol{\mu}_{t-1}, \quad (1)$$

where $\boldsymbol{\mu}_1 = \mathbf{x}_1$. It is ideal for capturing long term effects where its “memory” can be tuned using λ . Other options for capturing expected behavior exist as well, but the EWMA is easy to compute in a streaming setting.

We capture the deviation of observed behavior from expected behavior using residuals, defined as:

$$\mathbf{r}_t = \mathbf{x}_t - \boldsymbol{\mu}_t \quad (2)$$

Residuals are useful in identifying peculiarities such as bias or changing variance (heteroscedasticity). Note that our definition of a residual differs by a scaling factor from the conventional “look-ahead” residual defined for EWMA.

In univariate cases, residuals such as (2) are enough for outlier detection. In higher dimensions, however, (2) needs to be replaced by a careful choice of its multivariate analog. Distance metrics such as Mahalanobis distance are typically used to measure the deviation of a data point from its expected behavior. Most of the extant methods, however, are motivated by symmetric distributions such as the Gaussian and rely on a sphere of fixed radius to identify outliers, which might not be effective for skewed distributions.

We address this problem by defining a *signed distance*. We first use the Cholesky decomposition of $\hat{\Sigma}_0$, a robust variance-covariance estimate, to regularize the residuals,

$$\mathbf{e}_t = \hat{\Sigma}_0^{-1/2} \cdot \mathbf{r}_t. \quad (3)$$

The dominant component $p(t)$ is then decided by

$$p_t = \operatorname{argmax}_{j=1, \dots, d} |e_{tj}|. \quad (4)$$

We define the sample *signed Mahalanobis distance* of \mathbf{x}_t from its expected behavior as

$$d_t = \operatorname{sign}(e_{tp_t}) \cdot (\mathbf{x}_t - \boldsymbol{\mu}_t)^T \hat{\Sigma}_0^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_t). \quad (5)$$

Note that replacing the dispersion matrix Σ_0 with other types of matrices yields different types of distances. An identity matrix I_d yields the Euclidean distance, while a diagonal matrix Σ with diagonal elements $\sigma_{ii} = \text{Var}(x_{ti})$ makes d_t a standardized distance.

The signed distance (5) provides greater granularity, and hence greater accuracy in detecting outliers. It allows different thresholds for different directions of the data. Figure 1 illustrates a bivariate example. For distributions that are not symmetric, staggered boundaries shown in solid red and blue lines are more effective than the fixed dashed purple circle defined by an unsigned distance.

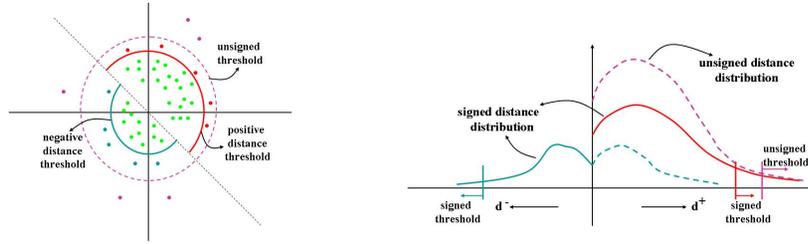


FIGURE 1. Asymmetric thresholds based on the signed distances.

Now let $q_t(\alpha)$ denote the α th quantile of the distribution of $\{d_t\}$. We define the set of outliers Ω as

$$\Omega = \{\mathbf{x}_t : d_t < q_t(\alpha) \text{ or } d_t > q_t(1 - \alpha)\}, \quad (6)$$

and keep track of the number of observed outliers using an *Outlier Curve*:

$$OC_t = \sum_{i=1}^t I_{\{\mathbf{x}_i \in \Omega\}}. \quad (7)$$

Since the thresholds $q_t(1 - \alpha)$ and $q_t(\alpha)$ are quantile based, we expect to detect a certain number of outliers by time t , given by

$$EOC_t = \mathbb{E}[OC_t] = \sum_{i=1}^t [P(d_i < q_i(\alpha)) + P(d_i > q_i(1 - \alpha))] = 2\alpha \cdot t, \quad (8)$$

which we call the *Expected Outlier Curve* at time t .

If there is a distributional change in the data stream, the quantile thresholds are no longer appropriate, and OC_t may deviate from EOC_t . The deviation is statistically significant if it violates the confidence interval for EOC_t of confidence level $100(1 - \beta)\%$,

$$\frac{OC(t) + z_{1-\beta/2}^2/2 \pm z_{1-\beta/2} \cdot \sqrt{OC(t)(1 - OC(t))/t + z_{1-\beta/2}^2/4t^2}}{1 + z_{1-\beta/2}^2/t}, \quad (9)$$

computed by noting that the average number of outliers, OC_t/t , is simply a sample proportion [Rao].

We hold $q_t(\alpha)$ and $q_t(1 - \alpha)$ fixed until $2\alpha \cdot t$ is outside the confidence interval defined in (9), when the null hypothesis that the data stream is stable does not hold. We then update these thresholds based on the current set of signed distances. To implement our method in a streaming fashion, we employ an efficient streaming algorithm for computing the quantiles [Greenwald *et al.*]. This quantiling is also accomplished in an adaptive fashion.

Our algorithm has many advantages. It is a real-time algorithm that can be used online. We update the summary statistics, including “mean”, sample variance-covariance matrix and quantiles, in a streaming manner. Our algorithm can detect a variety of distributional changes. For instance, in many cases, the dispersion of the data streams may decrease at some point. As a result, the number of observed outliers will decrease. Methods that rely solely on identifying “outliers” will miss this distribution change since “no outliers” is a normal status, while in our algorithm, OC_t will fall significantly below EOC_t , resulting in a re-calibration of the thresholds to reflect the new, tighter distribution. In addition, by introducing the concept of signed distances, we can handle skewed distributions without data transformations.

3 Applications

3.1 File Descriptor Streams

Our first application pertains to a data stream gathered from a file monitoring process. The calls made on a telecommunications network are logged and written to files in a highly specialized format. The files are gathered and processed for billing and network performance measurement purposes. A telecommunication company typically receives tens of thousands of such data files a day. There is usually a very small window of time to request a re-transmission of these files in case they are lost or damaged. Anomalies are acted upon immediately before the window of opportunity for retransmission is closed. An unusually large file size indicates data corruption or duplication, while smaller file sizes might indicate missing or damaged data. Lost or damaged files imply a loss in billing revenue.

A file monitoring process tracks the file stream by gathering file descriptors such as number of files received, their sizes and other characteristics, at short intervals of times. For the purposes of this paper, we track the sizes of three important file types across 3058 instances.

The left panel in Figure 2 shows the file sizes for the three file types polled at frequent time intervals and OC_t based on the three variables. Barring the initial period, the outlier curve shows a jump at around $t = 700$ and

a bigger jump at $t = 1,700$. EOC_t stays within the gray confidence bands until around $t = 1,800$. The confidence bands turns back into a safe (gray) mode thereafter, indicating that the distribution is stabilizing.

The file descriptor data stream is well behaved as expected, since it is generated by a well designed, well established data gathering process, except for occasional hiccups caused by maintenance. We were able to identify small changes that the processing center was not aware of, resulting in improvements to their work flow, as well as preventing significant revenue losses due to billing errors.

3.2 Server Usage Streams

The second application concerns monitoring a cluster of servers that supports an important e-commerce application. The application is critical enough that the servers need to be up all the time. The server usage data stream is used to monitor the health of the server cluster. For the purpose of this paper, the stream consists of four variables and 317,980 data points. However, the original stream has over 25 variables and accumulates in real-time.

Shown in the right panel of Figure 2 are the individual variables with the multivariate outliers denoted by magenta points. The outlier curve shows a small step-up around $t = 8,000$, consistent with the anomalies in variables B and C. We observe a big jump in OC_t around $t = 13,000$, when variables A and C experience a sudden level shift. The confidence bands are shown in pink when the confidence bounds are being re-calibrated when the distribution is unstable and changing. The two curves converge when the stream stabilizes and the bounds become stable and are shown in gray.

The engineers who monitor this application need to track just one curve, OC_t . They do not need to reconcile multiple univariate tests. The engineers can also control the number of deviants by choosing β appropriately. They decide which deviant to act upon based on their domain knowledge. Our method provides engineers with a fast, easy to use, and highly interpretable tool that saves time and avoids the overhead of multiple univariate tests.

4 Conclusions

We have presented an efficient streaming method for detecting distributional changes and characterizing these changes in multi-dimensional data streams. We have adapted techniques from several disciplines such as time series, nonparametric statistics and streaming algorithms for this purpose. The outlier curve is a key component that identifies and characterizes changes in the data stream by tracking the number of outliers. The shape of the curve reveals the natures of the changes in distribution. We have shown its effectiveness in real life applications.

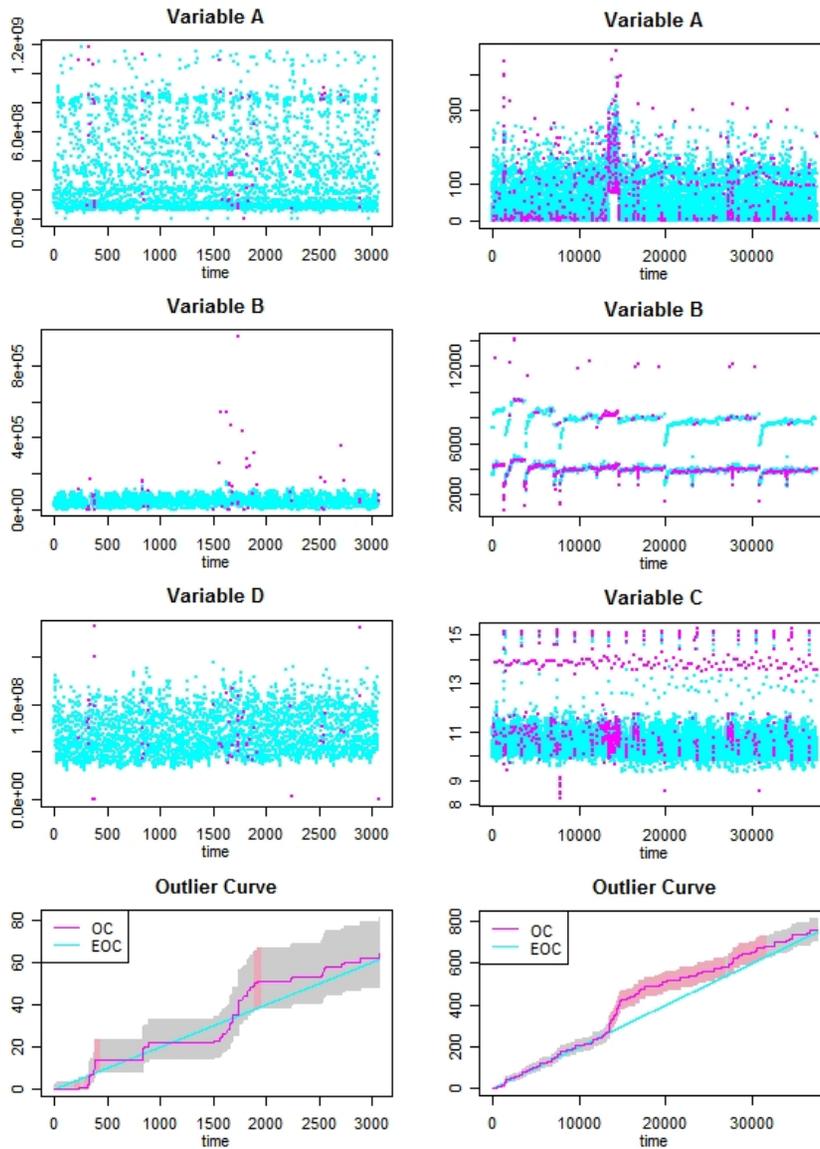


FIGURE 2. *Left:* File Descriptor Data Stream: Our outlier detection algorithm identifies anomalies in the data collection process that were previously not known to the processing center. Fixing the problem prevented significant revenue loss. *Right:* Server Usage Stream: Engineers use the outlier curve to track the distributional changes in the data stream.

References

- Babcock, B., Babu, S., Datar, M., Motwani, R. and Widom, J. (2002). Models and Issues in Data Stream Systems. In *Proceedings of PODS*.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B*, **57**, 125-133.
- Donoho, D. (1982). Breakdown Properties of Multivariate Location Estimators. Harvard University.
- Filzmoser, P., Maronna, R. and Werner, M (2008). Outlier Detection in High Dimensions. *Computational Statistics and Data Analysis*, **52**, 1694-1711.
- Gaber, M.M., Zaslavsky, A. and Krishnaswamy, S. (2005). Mining Data Streams: A Review. In *ACM SIGMOD Record*, **34**, 2.
- Galeano, P., Pena, D. and Tsay, R. S. (2006). Outlier Detection in Multivariate Time Series by Projection Pursuit. *Journal of American Statistical Association*, **101**, 474, 654-669.
- Greenwald, M. and Khanna, S. (2001). Space-efficient Online Computation of Quantile Summaries. In *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, 58-66.
- Hossjer, O. and Croux, C. (1995). Generalizing Univariate Signed Rank Statistics for Testing and Estimating a Multivariate Location Parameter. *Journal of Nonparametric Statistics*, **4**, 293-308.
- Huber, P. (1985). Projection Pursuit. *Annals of Statistics*, **13**, 2, 435-475.
- Liu, R., Singh, K. and Teng, J. (2004). DDMA-charts: Nonparametric Multivariate Moving Average Control Charts Based on Data Depth. *Advances in Statistical Analysis*, **88**, 235-258.
- Maronna, R. and Zamar, R (2002). Robust Estimates of Location and Dispersion for High-dimensional Data Sets. *Technometrics*, **44**, 4, 307-317.
- Pena, D. and Prieto, F. (2001). Multivariate Outlier Detection and Robust Covariance Matrix Estimation. *Technometrics*, **43**, 3, 286-310.
- Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*, Wiley.
- Rousseeuw, P.J. and Driessen, K.V. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, **41**, 3, 212-223.
- Rousseeuw, P. J. and Hubert, M. (1999). Regression Depth. *Journal of the American Statistical Association*, **94**, 388-402.

Space -Time Clustering Revisited

Gilbert MacKenzie¹ and Jing Xu¹

¹ Centre of Biostatistics, Department of Mathematics & Statistics, University of Limerick, Ireland

Abstract: The history of space-time clustering concepts and methods are reviewed briefly. The space-time clustering model of Ederer *et al* is investigated in detail. This method has been used extensively in the epidemiological literature, but we show that the distribution of main test statistic involved does not follow the distribution proposed by the authors. We note, too, that the two indices proposed are not statistically independent, leading to potential over-reporting in the epidemiological literature. We obtain the correlation between the original clustering indices and suggest a new combined test statistic which has the correct null distribution. We suggest how to develop a fuller spatial model and illustrate our methodology using data from a study of the incidence of childhood leukaemia in Northern Ireland.

Keywords: space-time clustering, exact distribution, correlated outcomes, spatial data, covariance model

1 Introduction

We review, briefly, the history of space-time clustering concepts and methods, ranging from Karl Pearson's now celebrated 1913 paper to modern times.

In particular we focus on a space-time clustering technique first introduced by Ederer *et al* who studied, *inter alia*, the spatial-temporal distribution of childhood leukaemia. Typically, this method, which has been used extensively in the epidemiological literature, relies on the creation of two indices of clustering Y_1 , the maximum number of observed cases in any single basic space-time unit, and Y_2 the maximum number of observed cases in any two time-contiguous basic units.

First, we demonstrate that that distribution of Y_1 , on the null hypothesis of no clustering, does not follow the null distribution proposed by the original authors. Second, we note that the two indices are not statistically independent. Consequentially, reports in the epidemiological literature to date may have been subject to over reporting. Third, using the exact distribution we compute the correlation between Y_1 and Y_2 for important cases which have appeared in the literature and discuss the problem of selecting a suitable bivariate test statistic. A new index which is asymptotically distributed as χ_2^2 is proposed. Its distribution in small samples is investigated by means of

a simulation study and a table of critical values is provided. Lastly, we consider how to develop a spatial model for the observed bivariate discrepancy in Y_1 & Y_2 from the null distribution. The new methodology is illustrated using data from the Northern Ireland Study of Patterns of Disease and Radiation (PODAR, 1989).

2 Basic Model Formulation

2.1 Space-Time Units

The basic idea is to detect unusual clusters of cases of a disease in space and time. A cluster occurs when some cases are distributed more closely in space and time than expected on the basis of some chance rule. The method of Ederer *et al* requires the creation of space-time units. A basic space-time unit is a defined spatial region studied for a unit of time. The individual spatial regions are referred to as spatial units and the units of time are referred to as temporal units. A study space-time unit is constructed by studying a spatial unit for several units of time, eg a region for say five years. It is assumed that the number of cases of disease can be determined in each year. Accordingly, let m = the total number of spatial units, k = the total number of temporal units and $N = m \times k$ be the total number of basic space-time units. With these arrangements we shall have m study space-time units each studied for k years. Ashitey & MacKenzie (1970) present a simple application.

2.2 Clustering Indices

Now, let n_{ij} be the number of cases of disease in the the i th spatial unit in temporal unit j : $i = 1, \dots, m$ and $j = 1, \dots, k$. And finally let $n_{i+} = \sum_{j=1}^k n_{ij}$ be the marginal total number of cases for the i th space-time unit, whence we have an occupancy distribution $(n_{i1} | n_{i,2}, \dots, | n_{i,k})$ with exactly k compartments. Ederer *et al* define two observed indices of clustering within the i th space-time unit, viz:

$$\begin{aligned} y_{i1} &= \text{maximum no. of cases in any 1 temporal unit} \\ y_{i2} &= \text{maximum no. of case in any 2 adjacent temporal units} \end{aligned} \quad (1)$$

These are just two possibilities.

2.3 Probability Functions

The distribution of the corresponding random variables, (Y_{i1}, Y_{i2}) and moments can be found in principle from:

$$Pr(n_{i1}, n_{i,2}, \dots, n_{i,k}) = n_{i+}! \prod_j \theta_j^{n_{ij}} / \prod_j n_{ij}! \quad (2)$$

where θ_j probability of falling into the j th. compartment. On the null hypothesis of no clustering $\theta_j = \theta = 1/k, \forall j$ whence we have an *occupancy* distribution problem (Feller, 1957) and, of course, there are many potential generalizations.

Given $Pr(n_{i1}, n_{i2}, \dots, n_{i,k})$, all possible sequences $\{n_{i1}, n_{i2}, \dots, n_{i,k}\}$ such that $\sum_{j=1}^k n_{ij} = n_{i+}$ may be found by exact enumeration, using an algorithm due to Nijenhuis Wilf (1978), whence the exact moments of the derived random variables (Y_{i1}, Y_{i2}) may be determined numerically. When exact numeration becomes time-consuming (typically for large k) simulation may be invoked.

3 Test Statistics

Ederer *et al* also define two continuity - corrected test statistics based on their indices, viz:

$$\begin{aligned} U_1^2 &= [|\sum_i Y_{i1} - E(\sum_i Y_{i1})| - 0.5]^2 / V(\sum_i Y_{i1}) \\ U_2^2 &= [|\sum_i Y_{i2} - E(\sum_i Y_{i2})| - 0.5]^2 / V(\sum_i Y_{i2}) \end{aligned} \quad (3)$$

and claim that they are distributed asymptotically as χ_1^2 random variables. We show, via an extensive simulation study, that this claim, in relation to U_1^2 , the most extensively quoted index in the epidemiological literature, is untenable.

Below we propose relatively simple solutions based on a test statistics which respects the bivariate distribution of (Y_{i1}, Y_{i2}) .

4 Presumed Independence

To date, we note, that routine usage of (U_1^2, U_2^2) has tacitly assumed that $Pr[(Y_{i1} \cap Y_{i2})] = Pr(Y_{i1}) \times Pr(Y_{i2})$. However, we shall establish that (Y_{i1}, Y_{i2}) is approximately BVN(μ, V) and that V is not a diagonal. In particular, below, we present in Table 1, the hitherto unpublished correlation between (Y_{i1}, Y_{i2}) , for a range of marginal case numbers $(y_{i,+})$ and different numbers of temporal units (k).

Broadly speaking, the correlation is non-trivial and this suggests that it should be taken into account when conducting tests using (U_1^2, U_2^2) .

Table 1: Exact Correlation between Y_{i1} and Y_{i2} :
for various k and y_{i+} (for arbitrary i)

y_{i+}	k=3	4	5	6	7	8	9
2	0.38	0.44	0.48	0.50	0.51	0.52	0.53
3	0.61	0.61	0.60	0.58	0.57	0.57	0.56
4	0.47	0.61	0.64	0.64	0.63	0.63	0.62
5	0.51	0.57	0.61	0.61	0.65	0.65	0.65
6	0.52	0.58	0.60	0.60	0.64	0.65	0.65

5 Analysis

5.1 Test Statistics

First we considered repairing the univariate random variable Y_{i1} by means of some simple transformations such as the square root and logarithmic in order to improve the tail behaviour, towards a χ_1^2 . In general, however, these transformations were not wholly successful whether they were applied to the sums over the study-spatial units or first at a study-spatial unit level.

Later, in the same spirit, we moved to investigate simple functions of Y_{i1} and Y_{i2} such the ratio $R_y = Y_{i1}/Y_{i2}$ and $\log_e(R_y)$. The results obtained here, while interesting, were rather *ad-hoc* and we gloss over them in this paper.

The move to a bivariate representation of (Y_{i1}, Y_{i2}) is based on a test chi-squared statistic of the general form

$$\chi_2^2 = [y - E(Y)]' V^{-1} [y - E(Y)] \quad (4)$$

Here, $[y - E(y)]' = [y_{+1} - E(y_{+1}), y_{+2} - E(y_{+2})]$ and $V = \Sigma V_i$ for $i = 1, \dots, m$. This was the most successful approach in a distributional sense as, in the simulation study, the test statistic was found to follow a chi-squared distribution with $\nu = 2$ degrees of freedom.

5.2 Residuals

The move to treating the indices of clustering jointly also opens up the possibility of defining residuals, at the spatial study level. These are useful for exploring departures from the null hypothesis of no clustering. With

these arrangements we shall have a pair of residuals $e'_i = (e_{i1}, e_{i2})$ for each study-spatial unit, ie, for $i = 1, \dots, m$. Thus, $x_i^2 = e'_i V^{-1} e_i$ will follow Unit Exponential distribution (approximately). Then the $x_{(i)}^2$ can be plotted against exponential order statistics to examine, in detail, any spatial discrepancy, thereby extending, considerably, the scope of Ederer's original scheme.

5.3 Simulation Study

Detailed simulation studies were carried out to support the conclusions reached in the analysis section. The truncated Poisson distribution (TPD) was used as a generating distribution for the marginal counts in the study-spatial units as necessarily $n_{i+} > 1 \quad i = 1, \dots, m$ and the TPD provides a convenient vehicle for simulating spatial dependence. More details will be given at the Workshop.

6 Data Analysis

Data from The Patterns of Disease with possible Association with Radiation, (PODAR) study, will be used to illustrate old and new methodologies. This study investigates the spatial pattern of the incidence of Childhood Leukaemia over a nine year period 1977-1985. It was commissioned as result of genuine public concern about the discharge of radioactive waste from the nuclear re-processing plant at Sellafield (Seascale, formerly Windscale, Cumbria, Lake District, England, UK) into the Irish Sea. It was conducted in Northern Ireland, UK, between 1987 and 1989. Some, $m = 526$ electoral wards comprised the spatial units. They were each studied for $k = 9$ years, leading to 526 study space-time units.

For example, during 1977-1985, in the city of Belfast there were 51 electoral wards, 35 of which had 2 or more cases of incident childhood leukaemia. In total there were 338 cases: $y_{+,1} = 172$, $E[y_{+,1}] = 168.48$ and $V[y_{+,1}] = 21.25$, leading to non-significant result in the city. More findings will be presented at the Workshop.

7 Final Remarks

This work in progress. We have uncovered some flaws in the original techniques proposed by Ederer *et al.* We have also been able to suggest some potential remedies as a result of our better understanding of the dependence between the two clustering indices gained primarily through direct computation and simulation. This latest latest approach opens up other avenues of evaluation and should lead to a more comprehensive spatial modelling approach. However, it is by no means complete.

Acknowledgments: This work was supported by Science Foundation Ireland (SFI) as part of their funded BIO-SI programme (www.ul.ie/bio-si), Mathematics Initiative, II. Dr. Jing Xu is a SFI funded Post-Doctoral Research Fellow

References

- Pearson K. (1913). Multiple cases of disease in the same house. *Biometrika*. **9**, 28-33.
- Ederer F., Myers MH. & Mantel, N. (1964). A statistical problem in space and time: do leukaemia cases come in clusters? *Biometrics*. **20**, 626.
- Ashity G. & MacKenzie G. (1970). 'Clustering' of multiple sclerosis cases by date and place of birth. *Brit. J. Prev. Soc. Medicine*. **24**, 163.
- Feller W. (1957) *An Introduction to Probability Theory and its Applications*. Wiley, New York.
- Nijenhuis A. & Wilf HS. *Combinatorial Algorithms* (2nd Edition), *Academic Press*, London 1978.

Confidence Intervals for Generalized Additive Model Components

Giampiero Marra¹ and Simon N. Wood²

¹ Corresponding author: Mathematical Sciences, University of Bath, Bath BA2 7AY, U.K. Email: g.marra@bath.ac.uk.

² Mathematical Sciences, University of Bath, Bath BA2 7AY, U.K.

Abstract: We study Bayesian confidence intervals for the smooth component functions of generalized additive models represented using any penalized regression spline approach. The intervals are the usual generalization of Wahba (1983) or Silverman (1985) intervals to the GAM component context. We present simulation evidence showing these intervals have close to nominal ‘across-the-function’ frequentist coverage probabilities, except when the truth is close to a straight line/plane function. We extend Nychka’s (1988) argument for univariate smoothing splines to explain these results. The theoretical argument suggests that good coverage probabilities can be achieved, provided that heavy oversmoothing is avoided, so that the bias is not too large a proportion of the sampling variability. Otherwise, because the Bayesian intervals account for bias and variance, the coverage probabilities are surprisingly insensitive to the exact choice of smoothing parameter. The theoretical results allow us to derive new intervals when a frequentist approach is taken, and to explain the impact that the neglect of smoothing parameter variability has on confidence interval performance. Also, they suggest switching the target of inference for component-wise intervals away from smooth components in the space of the GAM identifiability constraints. Instead intervals should be produced for each function as if only the other model terms were subject to identifiability constraints. If this is done then coverage probabilities are improved.

Keywords: Bayesian confidence interval; Demmler-Reinsch type parameterization; Generalized additive model (GAM); Penalized regression spline.

1 Introduction

A generalized additive model (GAM) can be seen as a generalized linear model (GLM) with a linear predictor dependent on smooth functions of covariates:

$$g\{\mathbb{E}(Y_i)\} = \mathbf{X}_i^* \boldsymbol{\theta}^* + \sum_j f_j(x_{ji}), \quad (1)$$

where $g(\cdot)$ is a link function, Y_i is a univariate response that follows an exponential family distribution, \mathbf{X}_i^* is the i th row of \mathbf{X}^* , which is the model

matrix for any strictly parametric model components, with corresponding parameter vector $\boldsymbol{\theta}^*$, and the f_j are smooth functions of the covariates \mathbf{x}_j , which may be vector covariates. The f_j are subject to identifiability constraints, such as $\sum_i f_j(x_{ji}) = 0 \forall j$. In this paper we concentrate on the case in which the f_j are represented using regression spline type bases, with associated measures of function ‘wiggleness’ that can be expressed as quadratic forms in the basis coefficients. Model fitting is via penalized iteratively reweighted least squares (P-IRLS) with smoothing parameters, λ_j , chosen minimizing the GCV score or generalized AIC (Wood, 2006, 2008).

Inference for *univariate* spline models can be effectively achieved using the Bayesian confidence intervals proposed by Wahba (1983) or Silverman (1985). As theoretically shown by Nychka (1988), for the case of univariate models whose Bayesian intervals have close to constant width, a very interesting feature of these intervals is that they work well when evaluated by a frequentist criterion, provided coverage is measured ‘across-the-function’ rather than pointwise. Specifically, consider the additive model

$$Y_i = f(x_i) + \epsilon_i, \epsilon_i \sim N(0, \sigma^2),$$

where the ϵ_i are mutually independent. According to Nychka’s results, if the smoothing parameter is sufficiently reliably estimated (e.g. by GCV) that the bias in the estimates is a modest fraction of the mean squared error for $f(x)$, then the average coverage probability (ACP)

$$\text{ACP}(\alpha) = \frac{1}{n} \sum_{i=1}^n \Pr[f(x_i) \in BI_\alpha(x_i)]$$

is very close to the nominal level $1 - \alpha$, where $BI_\alpha(x)$ indicates the $(1 - \alpha)100\%$ Bayesian interval for $f(x)$ and α the significance level. This agreement occurs because the Wahba/Silverman intervals include both a bias and variance component. These intervals as well as their component-wise extensions have been derived when dealing with Gaussian and non-Gaussian data. For example, by working in terms of the random pseudo-data vector \mathbf{z} , it has been shown (Wood, 2006) that the large sample posterior distribution for the regression spline coefficients of a GAM is

$$\boldsymbol{\beta} | \mathbf{y}, \boldsymbol{\lambda}, \phi \sim N(\hat{\boldsymbol{\beta}}, \mathbf{V}_\beta) \quad (2)$$

where $\hat{\boldsymbol{\beta}}$ is the maximum penalized likelihood estimate of $\boldsymbol{\beta}$ which is of the form $(\mathbf{X}^\top \mathbf{W} \mathbf{X} + \mathbf{S})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{z}$, $\mathbf{V}_\beta = (\mathbf{X}^\top \mathbf{W} \mathbf{X} + \mathbf{S})^{-1} \phi$, ϕ is a response distribution scale parameter, \mathbf{W} is the diagonal weight matrix at convergence of the P-IRLS algorithm used for fitting, and $\mathbf{S} = \sum_j \lambda_j \mathbf{S}_j$. The \mathbf{S}_j are known matrices which measure the wiggleness of the smooth functions.

Wang and Wahba (1995) compared the Bayesian confidence intervals with those derived using several variations of the bootstrap approach. They

found that the bootstrap framework can yield intervals that are comparable to the Bayesian ones in terms of across-the-function coverage properties. However, they are computationally intensive, problem which is common to the fully Bayesian approach as well.

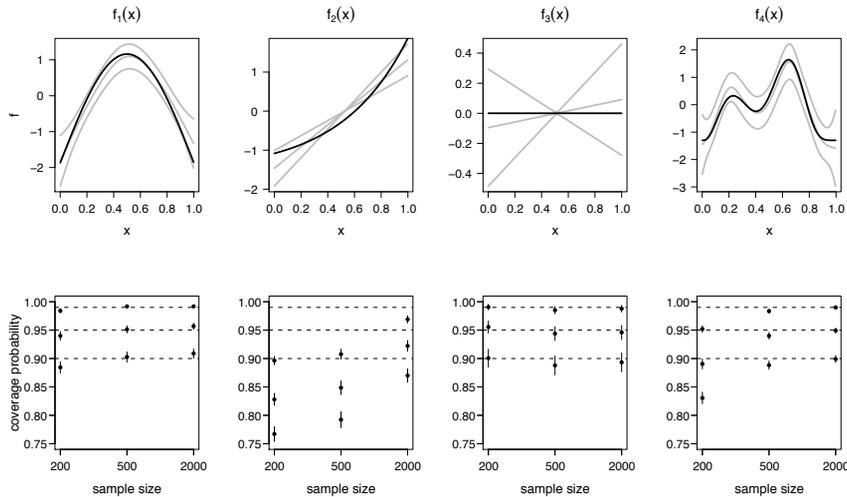


FIGURE 1. Results from component-wise Bayesian intervals for Bernoulli data. 1000 replicate datasets were generated and GAMs fitted using penalized thin plate regression splines (Wood, 2006, 2008). Top row: the true functions, indicated by the black lines, as well as typical estimates and 95% Bayesian confidence intervals (gray lines) for the smooths involved. \bullet represents the mean coverage probability, vertical lines show ± 2 standard error bands for the probabilities, and dashed horizontal lines the nominal probabilities considered.

Although Nychka's theoretical analysis has not been extended to non-constant width intervals for a smooth function that is a component of a larger model, simulation evidence suggests that result (2) might result in intervals for GAM components that perform well. As an example, Figure 1 illustrates that coverage probabilities for smooth functions with different degrees of complexity can be rather close to the nominal levels. However, it also suggests that, for smooth components almost completely in the penalty null space (such as $f_2(x)$), coverage probabilities can be too low. In this article we extend and modify Nychka's analysis to derive non-constant width intervals for GAM components. Our theoretical results show why non-constant Wahba/Silverman type intervals for smooth components work well in a frequentist setting, and explain when and why they fail. A fix for the case in which the intervals fail is suggested and new intervals are derived from a frequentist point of view. The impact that the neglect of smoothing parameter uncertainty has on confidence interval performance is also explained by our results.

2 Confidence intervals

2.1 Estimation of $\mathbb{E}\|\mathbf{B}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2/n$, and σ^2

The subsequent arguments will require that we can estimate the expected mean squared error of linear transformations of the coefficient estimates $\hat{\boldsymbol{\beta}}$, so this needs to be addressed first. We seek an estimate for the expected mean squared error, $\mathbb{E}(M_B)$, of $\mathbf{B}\hat{\boldsymbol{\beta}}$ where \mathbf{B} is a matrix of fixed coefficients. For the sake of space we summarize below the main results. From the Bayesian approach we have $\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\lambda}, \sigma^2 \sim N(\hat{\boldsymbol{\beta}}, \mathbf{V}_\beta)$, which implies

$$\mathbb{E}(M_B) = \frac{1}{n} \mathbb{E}\|\mathbf{B}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2 = \frac{1}{n} \text{tr}(\mathbf{B}^\top \mathbf{B} \mathbf{V}_\beta).$$

From a frequentist point of view

$$\mathbb{E}(M_B) = \frac{1}{n} \mathbb{E}\|\mathbf{B}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2 = \frac{1}{n} \text{tr}(\mathbf{B}^\top \mathbf{B} \mathbf{V}_{\hat{\boldsymbol{\beta}}}) + \frac{1}{n} \|\mathbf{B}(\mathbf{F} - \mathbf{I})\boldsymbol{\beta}\|^2 \quad (3)$$

where $\mathbf{F} = (\mathbf{X}^\top \mathbf{X} + \mathbf{S})^{-1} \mathbf{X}^\top \mathbf{X}$ and $\mathbf{V}_{\hat{\boldsymbol{\beta}}} = (\mathbf{X}^\top \mathbf{X} + \mathbf{S})^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X} + \mathbf{S})^{-1} \sigma^2$. The final term on the right hand side (the mean squared bias) can be estimated in a different way. By using a ‘natural’ Demmler-Reinsch type parameterization, we have that $\boldsymbol{\beta} \leftarrow \mathbf{U}^\top \mathbf{R} \boldsymbol{\beta}$ and $\mathbf{B} \leftarrow \mathbf{B} \mathbf{R}^{-1} \mathbf{U}$, hence $\mathbf{V}_\beta / \sigma^2 = \mathbf{F} = (\mathbf{I} + \mathbf{D})^{-1}$ and $\mathbf{V}_{\hat{\boldsymbol{\beta}}} = (\mathbf{I} + \mathbf{D})^{-2} \sigma^2$. We do not know $\boldsymbol{\beta}$, but if we knew the distribution of likely $\boldsymbol{\beta}$ values then we could estimate the bias term by its expectation according to that distribution. The natural assumption is to use the prior employed in the Bayesian analysis. It is then routine to show that

$$\frac{1}{n} \mathbb{E}_\beta \|\mathbf{B}(\mathbf{F} - \mathbf{I})\boldsymbol{\beta}\|^2 = \frac{1}{n} \text{tr}(\mathbf{B}^\top \mathbf{B} \mathbf{H}) \sigma^2$$

where \mathbf{H} is a diagonal matrix with elements $H_{ii} = D_i / (1 + D_i)^2$. Recognizing that $\mathbf{H} \sigma^2 = \mathbf{V}_\beta - \mathbf{V}_{\hat{\boldsymbol{\beta}}}$, we have the estimate

$$\widehat{\mathbb{E}(M_B)} = \frac{1}{n} \mathbb{E}\|\mathbf{B}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2 = \frac{1}{n} \text{tr}(\mathbf{B}^\top \mathbf{B} \mathbf{V}_\beta)$$

Reversing the re-parameterization confirms that this is simply

$$\widehat{\mathbb{E}(M_B)} = \frac{1}{n} \text{tr}(\mathbf{B}^\top \mathbf{B} (\mathbf{X}^\top \mathbf{X} + \mathbf{S})^{-1}) \sigma^2 \quad (4)$$

in the original parameterization.

If $\mathbf{B} = \mathbf{X}$ the result leads easily to the usual estimator

$$\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}{n - \text{tr}(\mathbf{F})}. \quad (5)$$

Alternatively, if we use the plug in estimate of the mean squared bias in (3), then we could use

$$\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 - \|\mathbf{X}(\mathbf{F}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})\|^2}{n - 2\text{tr}(\mathbf{F}) + \text{tr}(\mathbf{F}\mathbf{F}^\top)}. \quad (6)$$

2.2 Intervals

We now consider the construction of intervals for some component function of a model, such that $[f(x_1), f(x_2), \dots, f(x_n)]^T \equiv \mathbf{f} = \mathbf{X}\boldsymbol{\beta}$. Here \mathbf{X} may be the model matrix for just one model component: the matrix mapping the vector of all model coefficients to the evaluated values of just one smooth component (so, many of the columns of \mathbf{X} may be zero). The approach modifies Nychka's (1988) construction in order to obtain intervals of variable width, which are also applicable in the case where the function is only a component of a larger model.

Given some convenient constants, C_i , we seek a constant, D , such that

$$\text{ACP} = \frac{1}{n} \mathbb{E} \left\{ \sum_k \mathbb{I}(|\hat{f}(x_k) - f(x_k)| \leq z_{\alpha/2} D / \sqrt{C_i}) \right\} = 1 - \alpha$$

where 'ACP' denotes 'Average Coverage Probability', \mathbb{I} is an indicator function, α is a constant between 0 and 1 and $z_{\alpha/2}$ is the $\alpha/2$ critical point from a standard normal distribution. To this end, define $b(x) = \mathbb{E}\{\hat{f}(x)\} - f(x)$ and $v(x) = \hat{f}(x) - \mathbb{E}\{\hat{f}(x)\}$, so that $\hat{f} - f = b + v$. Defining I to be a random variable uniformly distributed on $\{1, 2, \dots, n\}$ we have

$$\begin{aligned} \text{ACP} &= \Pr \left(|b(x_I) + v(x_I)| \leq z_{\alpha/2} D / \sqrt{C_I} \right) \\ &= \Pr \left(|\sqrt{C_I} b(x_I) + \sqrt{C_I} v(x_I)| \leq z_{\alpha/2} D \right) \\ &= \Pr (|B + V| \leq z_{\alpha/2} D) \end{aligned}$$

where $B = \sqrt{C_I} b(x_I)$ and $V = \sqrt{C_I} v(x_I)$. We need to be able to approximate the distribution of $B + V$.

Let $[b(x_1), b(x_2), \dots, b(x_n)]^T \equiv \mathbf{b} = \mathbb{E}(\hat{\mathbf{f}}) - \mathbf{f}$. Hence, defining $\mathbf{c} = (\sqrt{C_1}, \sqrt{C_2}, \dots, \sqrt{C_n})^T$, we have

$$\mathbb{E}(B) = \sum_k \frac{1}{n} b(x_k) \sqrt{C_k} = \mathbf{c}^T (\mathbb{E}(\hat{\mathbf{f}}) - \mathbf{f}) / n.$$

In practice this quantity is very small (unless heavy oversmoothing is employed), but in any case it can be estimated as

$$\widehat{\mathbb{E}(B)} = \mathbf{c}^T (\hat{\mathbf{f}} - \hat{\mathbf{f}}) / n = \mathbf{c}^T \mathbf{X}(\mathbf{F}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) / n, \tag{7}$$

where $\mathbf{F} = (\mathbf{X}^T \mathbf{X} + \mathbf{S})^{-1} \mathbf{X}^T \mathbf{X}$.

Now consider V . Defining $\mathbf{v} = [v(x_1), v(x_2), \dots, v(x_n)]^T$, we have $\mathbb{E}(\mathbf{v}) = \mathbf{0}$, and hence

$$\mathbb{E}(V) = \sum_k \frac{1}{n} v(x_k) \sqrt{C_k} = 0.$$

The covariance matrix of \mathbf{v} is, $\mathbf{V}_{\hat{\mathbf{f}}} = \mathbf{X}\mathbf{V}_{\hat{\boldsymbol{\beta}}}\mathbf{X}^\top$, the same as that of $\hat{\mathbf{f}}$. Hence

$$\text{var}(V) = \sum_k \frac{1}{n} \mathbb{E}\{v(x_k)^2 C_k\} = \text{tr}(\mathbf{C}\mathbf{V}_{\hat{\mathbf{f}}})/n,$$

where \mathbf{C} is the diagonal matrix with leading diagonal elements C_i . Now since $\mathbf{v} \sim N(\mathbf{0}, \mathbf{V}_{\hat{\mathbf{f}}})$, V is a mixture of normals, which is inconvenient unless $[\mathbf{C}\mathbf{V}_{\hat{\mathbf{f}}}]_{ii}$ is independent of i . If this constant variance assumption holds then $V \sim N(0, \text{tr}(\mathbf{C}\mathbf{V}_{\hat{\mathbf{f}}})\sigma^2/n)$ (and the lack of dependence on i implies a lack of dependence on $b(x_i)$, implying independence of B and V).

It is the distribution of $V + B$ that is needed. $\mathbb{E}(V + B) = \mathbb{E}(B)$ and by construction $\text{var}(V + B) = \mathbb{E}(M) - \mathbb{E}(B)^2$ where

$$M = \frac{1}{n} \sum_k C_k \{ \hat{f}(x_i) - f(x_i) \}^2 = \|\sqrt{\mathbf{C}}(\hat{\mathbf{f}} - \mathbf{f})\|^2/n = \|\sqrt{\mathbf{C}}\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2/n.$$

Now we can exactly re-use Nychka's (1988) argument: provided B is small relative to V then $V + B$ will be approximately normally distributed, i.e. approximately

$$B + V \sim N(\mathbb{E}(B), \mathbb{E}(M) - \mathbb{E}(B)^2).$$

We can estimate $\mathbb{E}(B)$ and $\mathbb{E}(M)$ (by the results of the previous section).

So, defining $\hat{\sigma}_{bv}^2 = \widehat{\mathbb{E}(M)} - \widehat{\mathbb{E}(B)}^2$, we have the approximate result

$$B + V \sim N(\widehat{\mathbb{E}(B)}, \hat{\sigma}_{bv}^2).$$

Routine manipulation then results in

$$\hat{f}(x_i) - \widehat{\mathbb{E}(B)}/\sqrt{C_i} \pm z_{\alpha/2} \hat{\sigma}_{bv}/\sqrt{C_i} \tag{8}$$

as the definition of intervals achieving close to $1 - \alpha$ ACP (i.e. $D = \sigma_{bv}$). So, it is the fact that the convolution of B and V is close to a normal that leads the intervals to have good across-the-function coverage.

So far the choice of C_i has not been discussed, but the constant variance requirement, for $[\mathbf{C}\mathbf{V}_{\hat{\mathbf{f}}}]_{ii}$ to be independent of i , places strong restrictions on what is possible here. Two choices are interesting.

1. $C_i^{-1} = [\mathbf{V}_{\hat{\mathbf{f}}}]_{ii}$ ensures that the constant variance assumption is met exactly. Note that in this case, if we use (4) as the expected mean squared error estimate,

$$\widehat{\mathbb{E}(M)} = \frac{\hat{\sigma}^2}{n} \sum_i \frac{[\mathbf{X}\mathbf{V}_{\boldsymbol{\beta}}\mathbf{X}^\top]_{ii}}{[\mathbf{V}_{\hat{\mathbf{f}}}]_{ii}}.$$

In effect the resulting intervals are using the frequentist covariance matrix $\mathbf{V}_{\hat{\mathbf{f}}}$, but 'scaled up' to the 'size' of the Bayesian covariance matrix $\mathbf{V}_{\mathbf{f}} = \mathbf{X}\mathbf{V}_{\boldsymbol{\beta}}\mathbf{X}^\top$. Using the plug in estimator of (3) we have

$$\widehat{\mathbb{E}(M)} = 1 + \|\sqrt{\mathbf{C}}\mathbf{X}(\mathbf{F}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})\|^2/n \tag{9}$$

2. $C_i^{-1} = [\mathbf{V}_{\mathbf{f}}]_{ii}$. If $[\mathbf{V}_{\hat{\mathbf{f}}}]_{ii} \approx \gamma[\mathbf{V}_{\mathbf{f}}]_{ii}$ for some constant γ , then this choice approximately meets the constant variance assumption. If we use the (typically accurate) approximation $\widehat{\mathbb{E}(B)} \approx 0$, along with the mean squared error estimate (4), then the resulting intervals are exactly Bayesian intervals of the Wahba/Silverman kind.

In summary: we have shown that Bayesian, or our proposed alternative, intervals for the component smooth functions of an additive model should achieve close to nominal across the function coverage probability, provided only that we do not oversmooth so heavily that average bias dominates the sampling variability for a term estimate. Beyond this requirement not to oversmooth too heavily, the results appear to have rather weak dependence on smoothing parameter values, suggesting that the neglect of smoothing parameter variability should not significantly degrade interval performance. The results of this section can be routinely extended to the *generalized* additive model case.

Our results explain the success of ‘Bayesian’, component-wise, variable width intervals and the cases where the Bayesian intervals fail. The major failure, evident from simulations (see Figure 1), occurs when a smooth component is close to a function in the null space of the component’s penalty (i.e. to a straight line or plane) and may therefore be estimated as exactly such a function. The component will have been estimated subject to an identifiability constraint, but when intervals are constructed subject to such a constraint, the observed coverage probabilities are poor, and the preceding theory explains why: when a term is estimated as a straight line but *subject to an identifiability constraint* then the associated confidence interval necessarily has width 0 where the line passes through the zero line. In this case the sampling variability must be smaller than the bias over some interval surrounding the point, and the assumption that B is less than V will fail. The theory also suggests a remedy to this problem: compute each term’s interval as if it alone were unconstrained, and identifiability was obtained by constraints on the other model terms.

3 Some simulation results

Figure 2 shows some of the results obtained.

References

- Nychka, D. (1988). Bayesian Confidence Intervals for Smoothing Splines. *Journal of the American Statistical Association*, **83**, 1134-1143.
- Silverman, B. W. (1985). Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting. *Journal of the Royal Statistical Society Series B*, **47**, 1-52.

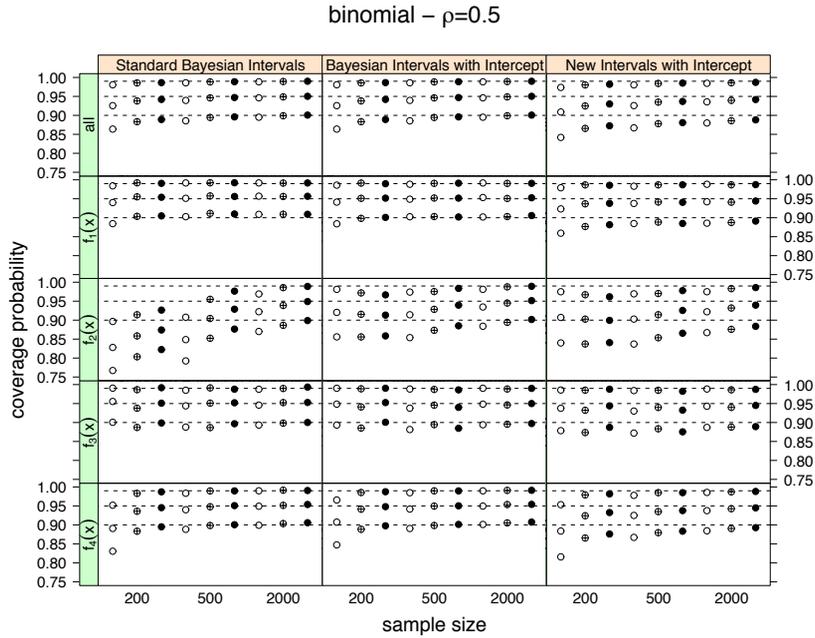


FIGURE 2. Coverage probability results for binomial data. Covariate correlation was equal to 0.5. \circ , \oplus and \bullet stand for high, medium and low noise level respectively. Standard error bands are not reported since they are smaller than the plotting symbols. Notice the improvement in the performance of the component-wise intervals for $f_2(x)$ when the fix is employed (see column 2).

Wahba, G. (1983). Bayesian ‘Confidence Intervals’ for the Cross-Validated Smoothing Spline. *Journal of the Royal Statistical Society Series B*, **45**, 133-150.

Wang, Y., and Wahba, G. (1995). Bootstrap Confidence Intervals for Smoothing Splines and Their Comparison to Bayesian Confidence Intervals. *Journal of Statistical Computation and Simulation*, **51**, 263-279.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. London: Chapman & Hall.

Wood, S. N. (2008). Fast Stable Direct Fitting and Smoothness Selection for Generalized Additive Models. *Journal of the Royal Statistical Society Series B*, **70**, 495-518.

Fitting DNA sequences through loglinear modeling with linear constraints

Nirian Martín¹ and Leandro Pardo²

¹ School of Statistics, Universidad Complutense de Madrid, Spain

² Mathematics Faculty, Universidad Complutense de Madrid, Spain

Abstract: For some discrete state series, such as DNA sequences, it can often be postulated that its probabilistic behaviour is given by a Markov chain. For making the decision on whether or not an uncharacterized piece of DNA is part of the coding region of a gene, under the Markovian assumption, there are two statistical tools that are essential to be considered: the hypothesis testing of the order in a Markov chain and the estimators of transition probabilities. In order to improve the traditional statistical procedures for both of them a new version for understanding the homogeneity hypothesis is proposed so that loglinear modeling is applied for conditional independence jointly with homogeneity restrictions on the expected means of transition counts in the sequence. In addition we can consider a variety of test-statistics and estimators by using power divergence measures. As special case of them the well-known likelihood ratio test-statistics and maximum likelihood estimators are obtained.

Keywords: Loglinear Model; Restricted Estimator; Conditional Test Statistic.

1 Introduction

The deoxyribonucleic acid (DNA) is composed by two linked sequences or chains of bases called adenine, cytosine, guanine, and thymine that are the four states of any DNA sequence. By knowing one of the two sequences the other one remains totally specified, for this reason when we are dealing with one sequence the other one is being omitted for its study. Coding adenine, cytosine, guanine, and thymine by 1, 2, 3 and 4 respectively, the DNA sequence of bases found within the first intron of the human preproglucagon gene of length 1572 is 34144111422341342423112411214...133131314441434434122114212434424421213 (the whole sequence is available from the web site www.ncbi.nlm.nih.gov of the National Center for Biotechnology Information, under the access number V01515, region 260..1831). The simplest Markovian model we can deal with is the so-called *Positional Independence* model. According to such a model, in a sequence of m random variables (r.v.) $\{X_i\}_{i=1}^m$, all of them in the same initial conditions, the result of X_i is independent of any of the preceding r.v.'s X_1, \dots, X_{i-1} . An equivalent model is a homogeneous Markov chain of order $p = 0$. The term

homogeneity come from the fact that the underlying r.v. is the same every time, $\Pr(X_i = x) = \Pr(X = x)$, $i = 1, \dots, m$. It is considered a discrete state serie because the support of X is discrete, $\mathcal{X} \equiv \{1, \dots, s\}$ with $s = 4$. A Markov chain is a discrete time serie because $i \in \mathcal{T} \equiv \{1, \dots, m\}$ indexes a period of time, however it is representing the position within the DNA sequence. Data from discrete state and discrete time series, $\{X_i\}_{i=1}^m$, which are modeled with *Markov chains of order* $p \in \mathbb{N}$ are formally defined so that the state of each individual depends on its states at the immediately preceding p instants but not on the rest of them, in other words, the future evolution of the process is conditionally independent of the past given the most recent p states

$$\begin{aligned} \Pr(X_{i+p} = h_{i+p} | X_{i'} = h_{i'}, i' \leq i + p - 1) = \\ \Pr(X_{i+p} = h_{i+p} | X_i = h_i, \dots, X_{i+p-1} = h_{i+p-1}), \quad i = 1, \dots, m - p. \end{aligned} \tag{1}$$

We are going to see the way in which the observations from a Markov chain of order p can be summarized in contingency tables. A contingency table is given by (N_1, \dots, N_s) , where $N_h = \sum_{i=1}^m \mathbb{I}(\{X_i = h\})$, $h = 1, \dots, s$ and $\mathbb{I}(A)$ is an indicator function (it is equal to 1 when it holds A and 0 otherwise), when $p = 0$. Observe that $n = \sum_{h=1}^s N_h$ is the contingency table's total number of observations. The multivariate r.v. (N_1, \dots, N_s) is a multinomial contingency table because its distribution is multinomial $\mathcal{M}(n; \pi_1, \dots, \pi_s)$, where $\pi_h = \Pr(X = h)$, $h = 1, \dots, s$. In a Markov chain of order p , the meaningful information is summarized in $m - p$ different $(p + 1)$ -way contingency tables

$$N_{h_1 \dots h_{p+1}}^{(i)} = \mathbb{I}(\{X_i = h_1, \dots, X_{i+p} = h_{p+1}\}), \quad i = 1, \dots, m - p. \tag{2}$$

If the probability distribution of the r.v. $X_{i+p} | X_i, \dots, X_{i+p-1}$ does not depend on i , i.e. $\Pr(X_{i+p} = h_{1+p} | X_{i+p-1} = h_p, \dots, X_i = h_1) = \Pr(X_{p+1} = h_{p+1} | X_p = h_p, \dots, X_1 = h_1)$, $i = 1, \dots, m - p$, then the Markov chain $\{X_i\}_{i=1}^m$, $j = 1, \dots, q$ is said to be homogeneous or stationary of order p . The advantage associated with a homogeneous Markov chain of order p (HMC(p)) is that all inferential procedures can be made by collapsing the $m - p$ contingency tables given in (2), i.e. we can deal with only one contingency table

$$N_{h_1 \dots h_{p+1}} \equiv \sum_{i=1}^{m-p} N_{h_1 \dots h_{p+1}}^{(i)} = \sum_{i=1}^{m-p} \mathbb{I}(\{X_i = h_1, \dots, X_{i+p} = h_{p+1}\}). \tag{3}$$

An interesting property held in any contingency table (3) of order $p \geq 1$ is

$$N_{h_1 \dots h_p * } \equiv \sum_{h=1}^s N_{h_1 h_2 \dots h_p h} = \sum_{h=1}^s N_{h h_2 \dots h_p h_{p+1}} \equiv N_{* h_1 \dots h_p}, \tag{4}$$

when the sequence does not start or end with the subsequence $h_1 \dots h_p$, or when the starting and the ending subsequence of length p is equal to $h_1 \dots h_p$, on the other hand

$$N_{h'_1 \dots h'_p * } = N_{* h'_1 \dots h'_p} + 1 \quad \text{and} \quad N_{h''_1 \dots h''_p * } = N_{* h''_1 \dots h''_p} - 1, \tag{5}$$

when the sequence does start with $h'_1 \cdots h'_p$ and end with $h''_1 \cdots h''_p$ being $h'_1 \cdots h'_p \neq h''_1 \cdots h''_p$.

When any state $h_{p+1} \in \mathcal{X}$ could be reached from any p -tuple $(h_p, \dots, h_1) \in \mathcal{X}^p$ of states, i.e. $\Pr(X_{i+p} = h_{1+p} | X_{i+p-1} = h_p, \dots, X_i = h_1) \equiv p_{h_1 \cdots h_{p+1}} > 0$, $i = 1, \dots, m - p$, the HMC(p) is said to be ergodic and the transition probability $p_{h_1 \cdots h_{p+1}}$ can be expressed as follows

$$p_{h_1 \cdots h_{p+1}} = \pi_{h_1 \cdots h_{p+1}} / \pi_{h_1 \cdots h_p}, \tag{6}$$

where $\pi_{h_1 \cdots h_p}$ is denoting the p -dimensional joint distribution

$$\Pr(X_i = h_1, \dots, X_{i+p-1} = h_p) \equiv \pi_{h_1 \cdots h_p} > 0, \quad i = 1, \dots, m - p. \tag{7}$$

Although $\{X_i = h_1, \dots, X_{i+p} = h_{p+1}\}_{i=1}^{m-p}$ is not a sequence of independent r.v.'s (unless $p = 0$), we can suppose that $(N_{h_1 \cdots h_{p+1}})_{(h_1, \dots, h_{p+1}) \in \mathcal{X}^{p+1}} \sim \mathcal{M}(n; (\pi_{h_1 \cdots h_{p+1}})_{(h_1, \dots, h_{p+1}) \in \mathcal{X}^{p+1}})$ with $n \equiv \sum_{(h_1, \dots, h_{p+1}) \in \mathcal{X}^{p+1}} N_{h_1 \cdots h_{p+1}} = m - (p + 1)$. Under multinomial sampling, we would like to test

$$\mathcal{H}_0(p): \{X_i\}_{i=1}^m \text{ is HMC}(p) \quad \text{vs.} \quad \mathcal{H}_1(p): \{X_i\}_{i=1}^m \text{ is HMC}(p + 1), \tag{8}$$

and if $\mathcal{H}_0(p)$ were accepted we would like to estimate its transition probabilities $p_{h_1 \cdots h_{p+1}}$.

For fitting the model under either $\mathcal{H}_0(p)$ or $\mathcal{H}_1(p)$, we are going to deal with the mean values $m_{h_1 \cdots h_{p+2}} \equiv E[N_{h_1 \cdots h_{p+2}}] = n\pi_{h_1 \cdots h_{p+2}}$ so that (6) is given by $p_{h_1 \cdots h_{p+1}} = m_{h_1 \cdots h_{p+1}*} / m_{h_1 \cdots h_p**}$. The classical maximum likelihood estimator (MLE) of $m_{h_1 \cdots h_{p+2}}$ under $\mathcal{H}_1(p)$ is the saturated model $\hat{m}_{h_1 \cdots h_{p+2}}^{\mathcal{H}_1(p)} = N_{h_1 \cdots h_{p+2}}$. Under $\mathcal{H}_0(p)$ we should consider that (1) can be expressed as

$$m_{h_1 h_2 \cdots h_{p+1} h_{p+2}} = m_{h_1 h_2 \cdots h_{p+1}*} \times m_{*h_2 \cdots h_{p+1} h_{p+2}} / m_{*h_2 \cdots h_{p+1}*}, \tag{9}$$

or equivalently as a conditional independent loglinear model (see Avery and Henderson (1999)), and thus the MLE is given by $\hat{m}_{h_1 h_2 \cdots h_{p+1} h_{p+2}}^{\mathcal{H}_0(p)} = N_{h_1 h_2 \cdots h_{p+1}*} \times N_{*h_2 \cdots h_{p+1} h_{p+2}} / N_{*h_2 \cdots h_{p+1}*}$. In order to follow the homogeneity assumption of order $p + 1$ more rigorously, it should be established

$$m_{h_1 \cdots h_{p+1}*} = m_{*h_1 \cdots h_{p+1}}, \quad \forall (h_1, \dots, h_{p+1}) \in \{1, \dots, s\}^{p+1} \tag{10}$$

under $\mathcal{H}_1(p)$ as well as under $\mathcal{H}_0(p)$, and this condition has not been considered in none paper until now. What is true is that taking into account $\hat{m}_{h_1 h_2 \cdots h_{p+1}*}^{\mathcal{H}_0(p)} = \hat{m}_{h_1 h_2 \cdots h_{p+1}*}^{\mathcal{H}_1(p)} = N_{h_1 h_2 \cdots h_{p+1}*}$, $\hat{m}_{*h_1 h_2 \cdots h_{p+1}}^{\mathcal{H}_0(p)} = \hat{m}_{*h_1 h_2 \cdots h_{p+1}}^{\mathcal{H}_1(p)} = N_{*h_1 h_2 \cdots h_{p+1}}$ and (4)-(5) for $p + 1$, the MLE's are not going to be in overall terms very different, however it seems that the accuracy of them and thus the quality of the test-statistics should be improved following this new idea. In the next section the manner of making statistical inferences through the model that consider not only (9) but also (10) is presented.

2 Statistical Inference through Loglinear Modeling with Linear Constraints (LMLC)

Let $\mathbf{m}(\boldsymbol{\theta}) = (m_1(\boldsymbol{\theta}), \dots, m_a(\boldsymbol{\theta}))^T$ be a vector of means $m_{h_1 h_2 \dots h_{p+1} h_{p+2}}(\boldsymbol{\theta})$ lexicographically ordered (in the following most of the vectors are denoted by a single index). Every loglinear model could be characterized by a full rank design matrix \mathbf{X} , such that $\log \mathbf{m}(\boldsymbol{\theta}) = \mathbf{X}\boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_b)^T$ is the vector of parameters of the loglinear model. Specifically for the conditional independence model ($\mathcal{H}_0(p)$) the dimensions of matrix \mathbf{X} are given by $a = s^{p+2}$ (number of rows) and $b = s^{p+2} - (s-1)^2 s^p$ (number of columns). Additionally we have to consider $c = s^{p+1} + 1$ constraints $\mathbf{L}^T \mathbf{m}(\boldsymbol{\theta}) = \mathbf{d}$, where the first one is $\mathbf{1}^T \mathbf{m}(\boldsymbol{\theta}) = n$ for the multinomial sampling constraint and the other ones are (10). Therefore \mathbf{L} is a $a \times r$ matrix, and $\mathbf{d} = (n, 0, \dots, 0)^T$. For $\mathcal{H}_1(p)$ the loglinear's design matrix is the identity matrix, $\tilde{\mathbf{X}} = \mathbf{I}_{s^{p+2}}$ ($\tilde{a} = \tilde{b} = s^{p+2}$), and the linear constraints are the same as for $\mathcal{H}_0(p)$, $\tilde{\mathbf{L}} = \mathbf{L}$, $\tilde{\mathbf{d}} = \mathbf{d}$ ($\tilde{c} = s^{p+1} + 1$). The new formulation of the hypothesis testing (8) in terms of LMLC is given by

$$\mathcal{H}_0(p) : M_0 \quad \text{vs.} \quad \mathcal{H}_1(p) : M_1, \tag{11}$$

where $M_0 = \{\mathbf{m}(\boldsymbol{\theta}) \in \mathbb{R}^a : \log \mathbf{m}(\boldsymbol{\theta}) = \mathbf{X}\boldsymbol{\theta}, \boldsymbol{\theta} \in \Theta_0\}$, $\Theta_0 = \{\boldsymbol{\theta} \in \mathbb{R}^{\tilde{b}} : \mathbf{L}^T \mathbf{m}(\boldsymbol{\theta}) = \mathbf{d}\}$ and $M_1 = \{\mathbf{m}(\boldsymbol{\theta}) \in \mathbb{R}^a : \log \mathbf{m}(\boldsymbol{\theta}) = \tilde{\mathbf{X}}\boldsymbol{\theta}, \boldsymbol{\theta} \in \Theta_1\}$, $\Theta_1 = \{\boldsymbol{\theta} \in \mathbb{R}^{\tilde{b}} : \tilde{\mathbf{L}}^T \mathbf{m}(\boldsymbol{\theta}) = \tilde{\mathbf{d}}\}$. Observe that $M_0 \subset M_1$, which means that model M_0 is nested within M_1 and this is why (8) or equivalently (11) are conditional hypothesis testing.

The MLE of $\mathbf{m}(\boldsymbol{\theta})$ under $\mathcal{H}_0(p)$ is given by $\mathbf{m}(\hat{\boldsymbol{\theta}}^{\mathcal{H}_0(p)})$ where

$$\hat{\boldsymbol{\theta}}^{\mathcal{H}_0(p)} = \arg \max_{\boldsymbol{\theta} \in \Theta_0} \ell(\mathbf{n}, \mathbf{m}(\boldsymbol{\theta})) + \boldsymbol{\mu}^T (\mathbf{L}^T \mathbf{m}(\boldsymbol{\theta}) - \mathbf{d}), \tag{12}$$

$\ell(\mathbf{n}, \mathbf{m}(\boldsymbol{\theta})) = \sum_{i=1}^a N_i \log m_i(\boldsymbol{\theta}) - \sum_{i=1}^a m_i(\boldsymbol{\theta})$ is the kernel of the loglikelihood function, $\mathbf{n} = (N_1(\boldsymbol{\theta}), \dots, N_a(\boldsymbol{\theta}))^T$ the vector of observed frequencies $N_{h_1 h_2 \dots h_{p+1} h_{p+2}}$ lexicographically ordered and $\boldsymbol{\mu} = (\mu_1, \dots, \mu_c)^T$ is the vector of Lagrange multipliers. Under $\mathcal{H}_1(p)$ estimator $\mathbf{m}(\hat{\boldsymbol{\theta}}^{\mathcal{H}_1(p)})$ is obtained in a similar manner replacing Θ_0 by Θ_1 . The likelihood ratio test-statistic is given by

$$G^2 = 2 \sum_{i=1}^a m_i(\hat{\boldsymbol{\theta}}^{\mathcal{H}_1(p)}) \log \frac{m_i(\hat{\boldsymbol{\theta}}^{\mathcal{H}_1(p)})}{m_i(\hat{\boldsymbol{\theta}}^{\mathcal{H}_0(p)})}, \tag{13}$$

whose asymptotic distribution under $\mathcal{H}_0(p)$ is going to be determined later. It is interesting to observe that (12) and (13) can be expressed in terms of the Kullback divergence measure between two non-negative vectors $\hat{\boldsymbol{\theta}}^{\mathcal{H}_0(p)} = \arg \min_{\boldsymbol{\theta} \in \Theta_0} D_{Kull}(\mathbf{n}, \mathbf{m}(\boldsymbol{\theta})) + \boldsymbol{\mu}^T (\mathbf{L}^T \mathbf{m}(\boldsymbol{\theta}) - \mathbf{d})$, $G^2 = 2D_{Kull}(\mathbf{m}(\hat{\boldsymbol{\theta}}^{\mathcal{H}_1(p)}), \mathbf{m}(\hat{\boldsymbol{\theta}}^{\mathcal{H}_0(p)}))$, where $D_{Kull}(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^a u_i \log(u_i/v_i) -$

$\sum_{i=1}^a u_i + \sum_{i=1}^a v_i$, $\mathbf{u} = (u_1, \dots, u_a)^T$, $\mathbf{v} = (v_1, \dots, v_a)^T$ (for more information about these measures see Martín and Pardo (2008)). The Kullback divergence measure is member of an important family of divergences called power divergence measures whose basis was established by Read and Cressie (1988) and it is defined for each value $\lambda \in \mathbb{R}$ as follows

$$D_{(\lambda)}(\mathbf{u}, \mathbf{v}) = \frac{2}{\lambda(1+\lambda)} \left(\sum_{i=1}^a \frac{u_i^{\lambda+1}}{v_i^\lambda} - \sum_{i=1}^a u_i + \lambda (\sum_{i=1}^a v_i - \sum_{i=1}^a u_i) \right), \tag{14}$$

for $\lambda \notin \{0, -1\}$, and $D_{(\lambda)}(\mathbf{u}, \mathbf{v}) = \lim_{s \rightarrow \lambda} D_{(s)}(\mathbf{u}, \mathbf{v})$, for $\lambda \in \{0, -1\}$. It can be checked that $D_{(0)}(\mathbf{u}, \mathbf{v}) = D_{Kull}(\mathbf{u}, \mathbf{v})$ and therefore on one hand (12) is a particular case the minimum power divergence estimators

$$\hat{\boldsymbol{\theta}}_{(\lambda)}^{\mathcal{H}_i(p)} = \arg \min_{\boldsymbol{\theta} \in \Theta_i} D_{(\lambda)}(\mathbf{n}, \mathbf{m}(\boldsymbol{\theta})) + \boldsymbol{\mu}^T (\mathbf{L}^T \mathbf{m}(\boldsymbol{\theta}) - \mathbf{d}), \quad i \in \{0, 1\}, \tag{15}$$

and on the other hand (13) is a special type of power divergence family of goodness of fit tests

$$T_{(\lambda, \lambda')} = 2D_{(\lambda')}(\mathbf{m}(\hat{\boldsymbol{\theta}}_{(\lambda)}^{\mathcal{H}_1(p)}), \mathbf{m}(\hat{\boldsymbol{\theta}}_{(\lambda)}^{\mathcal{H}_0(p)})). \tag{16}$$

Observe that $\hat{\boldsymbol{\theta}}_{(0)}^{\mathcal{H}_i(p)} = \hat{\boldsymbol{\theta}}^{\mathcal{H}_i(p)}$ is the MLE and the power divergence measure is applied twice in (16), one for estimation (λ) and other one for the test-statistic itself (λ'), in fact G^2 is obtained with $\lambda = \lambda' = 0$. The asymptotic distribution of (16) under $\mathcal{H}_0(p)$ is chi-squared with $(s - 1)^2 s^p$ degrees of freedom. This result is based on the next theorem (see Theorem 1 in Martín and Pardo (2006)), valid for testing (11) with any pair of nested LMLC.

Theorem 1. *Let $\mathcal{C}(\bullet)$ be the column space of a matrix. Suppose $\mathcal{C}(\mathbf{X}) \subset \mathcal{C}(\tilde{\mathbf{X}})$ or $\mathcal{C}(\tilde{\mathbf{L}}) \subset \mathcal{C}(\mathbf{L})$ and $\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{L}, \tilde{\mathbf{L}}$ are full rank matrices. The asymptotic distribution of the test-statistic $T_{(\lambda, \lambda')}$ for fitting M_0 given that it holds M_1 , is chi-squared with $\tilde{b} - b - \tilde{c} + c$ degrees of freedom.*

3 Simulation study

Our simulation study is carried out for a Markov chain of order $p = 1$, being the theoretical transition probabilities the same as given in Avery and Henderson (1999) for preproglucagon and the length of studied sequences are $m = \{600, 1500, 3000, 6000\}$. We fit the model for each of them with $R = 10,000$ replications. The mean squared error (MSE) of the estimators of the transition probabilities is $MSE(p) = \sum_{h_1 h_2} \sum_{h=1}^R (\tilde{p}_{h_1 h_2}(h) - p_{h_1 h_2})^2 / (s^{p+1} R)$, with $\tilde{p}_{h_1 h_2}(h) = m_{h_1 h_2}^{(h)}(\hat{\boldsymbol{\theta}}_{(\lambda)}^{\mathcal{H}_0(p)}) / m_{h_1}^{(h)}(\hat{\boldsymbol{\theta}}_{(\lambda)}^{\mathcal{H}_0(p)})$ for the new model and $\tilde{p}_{h_1 h_2}(h) = N_{h_1 h_2}^{(h)} / N_{h_1}^{(h)}$ for the traditional MLE's. We shall distinguish both of them by $MSE_{(\lambda)}(p)$ and $MSE_*(p)$ respectively. Regarding the new test-statistics $T_{(\lambda, \lambda')}$ and the classical likelihood ratio test-statistic T_* (see Avery and Henderson (1999)) we have evaluated

some characteristics such as for example how is the precision of the estimated expected value $\widehat{E}[T] = \sum_{h=1}^R T^{(h)}/R$ that should be next to 36 ($T \in \{T_{(\lambda, \lambda')}, T_{\star}\}$), or the exact value of the size when the significance level is $\alpha = 0.05$, $\widehat{\alpha} = \sum_{h=1}^R \mathbf{I}(\{T^{(h)} > \chi_{36, \alpha}^2\})/R$. Table 1 summarizes some results for $m = 3000$. For all values of m it has been seen that $MSE_{(0)}(p) < MSE_{\star}(p)$ and for all m except for $m = 600$, $MSE_{(1)}(p) < MSE_{(0)}(p)$. Therefore the new proposed estimators are more accurate than the classical ones and in addition the minimum chi-squared estimators ($\lambda = 1$) are even more accurate than the MLE's. The most precise test-statistic is obtained with $\lambda = 0$ and $\lambda' = 1$ and the likelihood ratio test-statistic, either the classical T_{\star} or the new one $T_{(0,0)}$, are quite imprecise.

TABLE 1. Simulation study of some accuracy properties of estimators and test statistics for $m = 3000$.

	λ	0	$\frac{2}{3}$	1	classic(★)		
$MSE(p) \times 10^4$		2.7167	2.7622	2.6814	2.7172		
(λ, λ')	(0, 0)	$(0, \frac{2}{3})$	(0, 1)	(1, 0)	$(1, \frac{2}{3})$	(1, 1)	classic(★)
$\widehat{E}[T]$	37.0622	36.0764	35.9769	37.5781	35.8336	35.3898	37.0616
$\widehat{\alpha}$	0.0660	0.0507	0.0502	0.0757	0.0468	0.0406	0.0667

4 Concluding remarks

We have presented estimators and test-statistics with better behavior than the traditional ones. As future research we expect that following the new proposed model and changing the way for collecting the data into the contingency table, the results could be considerably improved.

References

Avery PJ, Henderson DA (1999). Fitting Markov chain models to discrete state series such as DNA sequences. *J R Statist Soc C*, **48**, 53-61.

Martin N, Pardo L (2006). Phi-divergence tests statistics in multinomial sampling for hierarchical sequences of loglinear models with linear constraints. In: *Seminario Matemático Garcia Galdeano*, **31**, 301-308.

Martín N, Pardo L (2008). New families of estimators and test statistics in loglinear models. *J Multivariate Analysis*, **99**, 1590-1609.

Read TRC, Cressie NAC (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer.

A statistical model for the relation between exoplanets and their host stars

Elizabeth Martínez-Gómez^{1,2} and G. Jogesh Babu²

¹ Instituto de Astronomía, Universidad Nacional Autónoma de México, Apdo. Postal 70-264, Ciudad Universitaria, 04510, México D. F., México

² Center for Astrostatistics, 326 Thomas Building, The Pennsylvania State University, University Park, PA, 16802-2111, USA

Contact and presenting author: affabeca@gmail.com

Abstract: A general model is proposed to explain the relation between the extrasolar planets (or exoplanets) detected until June 2008 and the main characteristics of their host stars through statistical techniques. The main goal is to establish a mathematical relation among the set of variables which better describe the physical characteristics of the host star and the planet itself. The host star is characterized by its distance, age, effective temperature, mass, metallicity, radius and magnitude. The exoplanet is described through its physical parameters (radius and mass) and its orbital parameters (distance, period, eccentricity, inclination and major semiaxis). As a first approach we consider that only the mass of the exoplanet is being determined by the physical properties of its host star. The proposed model is then validated through statistical analysis.

Keywords: Linear regression, Cross-sectional data, Exoplanets

1 Introduction

An extrasolar planet (or exoplanet) is a planet which orbits a star other than the Sun, and therefore belongs to a planetary system other than our Solar System. The first extrasolar planet around a main sequence star was discovered in 1995 (Mayor and Queloz, 1995). Actually more than 300 exoplanets have been documented and most of them with masses greater than Jupiter's mass (Schneider, 2009). Detecting an exoplanet is a very difficult task because they do not emit any electromagnetic radiation of their own and are completely obscured by their extremely bright host stars, that is, normal telescope observation techniques cannot be used. Thus, in order to find exoplanets, a variety of techniques like the radial velocity, pulsar timing, astrometry, gravitational lensing, spectrometry and photometry (De Pater and Lissauer, 2001) are used. The main purpose of any method is to detect the effect produced by the exoplanet on its stellar system. Besides the discoveries it is important to search for models that can explain the origin, formation and possible migration of these bodies. For example,

Rice and Armitage (2005) have investigated how the statistical distribution of extrasolar planets may be combined with knowledge of the host stars' metallicity to yield constraints on the migration histories of gas giant planets. Moreover in a series of papers (Udry et al., 2003; Santos et al., 2003; Eggenberger et al., 2004; Halbwegs et al., 2005) the emerging properties of planet-host stars and characteristics of the different orbital-element distributions of exoplanetary systems have been studied. In this work we analyze the cross-sectional data for the exoplanets detected until June 2008 through linear regression techniques. The purpose of this kind of analysis is to verify the relation between the host star and its orbiting planet. For example, if the planet's mass is strongly determined by the type of star and hence affects the planetary formation stage.

1.1 Characteristics of the data catalog: Stars and Planets

The catalog was created in February 1995 to facilitate the progress of the new field named Exoplanetology through the publication of recent detections and their associated data. The catalog is interactive and it is available in the webpage: <http://exoplanet.eu>.

Until June 2008 the catalog contains: 303 exoplanets and 259 planetary systems (31 multiple systems). Two important considerations are: 1) the mass of the exoplanet is -at least- $13 M_J$ (Jupiter's mass) and 2) the data source must be reliable, that is, previously published in referred journals, presented in conferences, among others.

- **Stars:** The stellar data are taken from well-known databases like Simbad or directly from published papers. The basic physical characteristics of a star are: radial velocity, mass, metallicity, age and distance.
- **Planets:** These data are taken from published papers and from the sites: Anglo-Australian Planet Search; California and Carnegie Planet Search; Geneva Extrasolar Planet Search Programmes; Transatlantic Exoplanet Survey and the Department of Astronomy at University of Texas.

2 The General Model: Multiple Regression Analysis

We start with the following model (Model A) described by the equation:

$$M_P = \alpha_1 + \alpha_2 DS + \alpha_3 AS + \alpha_4 TS + \alpha_5 MS + \alpha_6 METAL + \alpha_7 MAG + \alpha_8 RS + u_i \quad (1)$$

where M_P is the exoplanet's mass and α_i are the coefficients for each term. Eq. (1) expresses the exoplanet's mass M_P in terms of the values of the

TABLE 1. Estimated values for the parameters.

Model A				
Variable	α_i	Standard Error	t-statistic	Probability
<i>C</i>	-9.1773	5.0051	-1.8336	0.0685
<i>DS</i>	-0.0113	0.0074	-1.5215	0.1301
<i>ES</i>	0.0288	0.0872	0.3299	0.7419
<i>TS</i>	0.0013	0.0007	1.9169	0.0570
<i>MS</i>	1.9385	1.3721	1.4128	0.1596
<i>METAL</i>	-1.7493	1.2586	-1.3899	0.1664
<i>MAG</i>	0.3689	0.2850	1.2944	0.1973
<i>RS</i>	0.2335	0.1183	1.9747	0.0499
Model B				
Variable	α_i	Standard Error	t-statistic	Probability
<i>C</i>	-2.5169	1.0840	-2.3218	0.0213
<i>ES</i>	-0.0493	0.0321	-1.5345	0.1265
<i>TS</i>	0.0003	0.0002	1.7417	0.0831
<i>MS</i>	1.1772	0.3964	2.9698	0.0033
<i>METAL</i>	-1.1370	0.5001	-2.2738	0.0241
<i>SIST*MS</i>	-0.1809	0.2188	-0.8269	0.4093
<i>SIST*METAL</i>	0.5629	0.8338	0.6752	0.5003

variables representing the features of the host star. This set of variables contains: the distance, *DS*; the age, *AS*; the temperature, *TS*; the mass, *MS*; the metallicity, *METAL*; the magnitude, *MAG* and the radius, *RS*. Finally u_i are the random errors.

We estimate the unknown parameters in Eq. (1) by Ordinary Least Squares (OLS). The results are shown in Table 1 where we also include the values of the t-statistics and their associated probabilities for the coefficient significance tests. From the estimated values we conclude that the only significant variable for the Model A is *RS*.

2.1 Verification of the linear regression assumptions (Model A)

1. Linearity: The model passed all the Ramsey tests for linearity. We conclude that the proposed functional form is adequate.
2. Omitted Variables: According to the star formation theory, the variables *MS* and *METAL* must be included to explain the relation between the mass of the exoplanet and its host star.
3. Multicollineality: There is a possible weak correlation between *MAG* and *DS*.

4. Heteroskedasticity: From the White test on the residuals we conclude that they are not heteroskedastic, that means the residuals are homoskedastic.
5. Normality: From the value of the Jarque-Bera statistic we conclude that the residuals are not normally distributed.
6. Homogeneity: Defining the "dummy" variable as *SIST* (0 means that the exoplanet belongs to a single planetary system and 1 refers to a multiple planetary system) we conclude that the model is homogeneous.

Statistical model must satisfy all the assumptions mentioned above to be correctly specified. In our case, the Model A needs some modifications, for example, another functional form and/or the consideration of an adequate "dummy" variable. In such a case we derive the Model B:

$$\log(M_P) = \alpha_1 + \alpha_2 ES + \alpha_3 TS + \alpha_4 MS + \alpha_5 METAL + \gamma_1 SMS + \gamma_2 SMET + u_i \quad (2)$$

where $SMS = SIST * MS$ and $SMET = SIST * METAL$ are two new variables that take into account the fact that the exoplanet can belong to a single or a multiple planetary system. The parameters are estimated through OLS and the results are summarized in Table 1.

2.2 Verification of the linear regression assumptions (Model B)

1. Linearity: The model passed all the Ramsey tests for linearity. Moreover we conclude that the new functional form is more adequate than the presented in Model A.
2. Omitted Variables: The tests indicate that the variables *ES* and *TS* must be excluded. However, under this situation the linearity is not preserved and we loose important physical information about the host star.
3. Multicollineality: There is no correlation among the independent variables.
4. Heteroskedasticity: From the White test on the residuals we conclude that they are heteroskedastic.
5. Normality: From the value of the Jarque-Bera statistic we conclude that the residuals are normally distributed.
6. Homogeneity: The model has already included the effect of a dummy variable.

Model B (*log-linear*) is slightly better than Model A in the sense that we have improved some of the discrepancies previously detected in the basic assumptions. However, this latter model cannot be considered yet to explain the relation between an exoplanet and its host star. Including the effect of a "dummy" variable seems to be a clue for another type of model. This binary behavior is discussed in the next section.

3 Multiple Regression Analysis with Binary Dependent Variables: a different approach

Based on the data, the dependent variable (exoplanet) is simultaneously determined by several parameters, qualitative and quantitative. In this work we have just assumed that the mass, M_P , is the quantitative variable that represents the whole physical/orbital characteristics of the planet. However, this fact is not completely true and more qualitative information must be taken into account for the model.

In the context of the variable *exoplanet*, the relevant information can be captured by defining a binary variable or a zero-one variable. An example of such a variable was introduced in Section 2 as *SIST* and it is related to the fact that the exoplanet can belong to a single or a multiple planetary system, in other words, $SIST = 0$ if the exoplanet belongs to a single planetary system and $SIST = 1$ in other case.

Under this new approach some binary models can be employed and their choice depends on the data distribution. For example, for a normal distribution we apply the *probit* model, for a logistic distribution we apply the *logit* model and when the data are truncated or censored we apply the *tobit* model.

Once the model is selected, its parameters can be estimated through the traditional methods like the Maximum Likelihood (ML) and Ordinary Least Squares (OLS).

A general binary model (Model C) for this case can take the form:

$$M_P = \alpha_1 + \alpha_2 ES + \alpha_3 TS + \alpha_4 MS + \alpha_5 METAL + u_i \quad (3)$$

The special case of Model C under the binary context will be discussed elsewhere. Recently a *logit* model was developed and validated by Fressin et al. (2009). In that work the authors performed a logistic regression to model the probability that a given planet is "real" (that means, observed or detected) or just simulated.

4 Summary and Conclusions

From our extensive statistical analysis we conclude that Model B is better than Model A. We have improved its specifications through the deletion of

variables like *MAG*, *DS* and *RS* and the addition of new ones that consider the possibility of finding exoplanets in single or multiple planetary systems. At the moment this is our best representation of the relation between the exoplanet and its host star and in a future work we will consider the problem by including binary variables.

Acknowledgements

E. Martínez-Gómez thanks to DGAPA-UNAM postdoctoral fellowship and to the Faculty for the Future Program (Schlumberger Foundation) for the financial support provided for this work. G. Jogesh Babu is supported in part by a National Science Foundation grant AST-0707833.

5 References

- De Pater, I. and Lissauer, J. J. (2001). *Planetary Sciences*, Cambridge University Press, Chapter 13, 576 pp.
- Eggenberger, A., Udry, S. and Mayor, M. (2004). Statistical properties of exoplanets. III. Planet properties and stellar multiplicity. *Astronomy and Astrophysics*, **417**, 353-360.
- Fressin, F., Guillot, T. and Nesta, L. (2009). Interpreting the yield of transit surveys: Are there groups in the known transiting planets population?. *Astronomy and Astrophysics*, in press.
- Halbwachs, J. L., Mayor, M. and Udry, S. (2005). Statistical properties of exoplanets. IV. The period-eccentricity relations of exoplanets and of binary stars. *Astronomy and Astrophysics*, **431**(3), 1129-1137.
- Mayor, M. and Queloz, D. (1995). A Jupiter-mass companion to a solar-type star. *Nature*, **378**, 355-359.
- Rice, W. K. M. and Armitage, P. J. (2005). Quantifying Orbital Migration from Exoplanet Statistics and Host Metallicities. *The Astrophysical Journal*, **630**(2), 1107-1113.
- Santos, N. C., Israelian, G., Mayor, M., Rebolo, R. and Udry, S. (2003). Statistical properties of exoplanets. II. Metallicity, orbital parameters, and space velocities. *Astronomy and Astrophysics*, **398**, 363-376.
- Schneider, J. (2009). Interactive Extra-Solar Planets Catalog. (<http://exoplanet.eu>).
- Udry, S., Mayor, M. and Santos, N. C. (2003). Statistical properties of exoplanets. I. The period distribution: Constraints for the migration scenario. *Astronomy and Astrophysics*, **407**, 369-376.
- Wooldridge, J. M. (2008). *Introductory Econometrics: A Modern Approach*, South-Western College Pub., 4th ed., 865 pp.

A clustering method for ordered variables to detect up-correlated genomic regions

Tristan Mary-Huard¹, Émilie Lebarbier¹ and Stéphane Robin¹

¹ AgroParisTech/INRA, UMR 518, Mathématiques et Informatique Appliquées, F-75005 Paris, France

Abstract: We propose a clustering based method to identify chromosomal domains of gene expression. The strategy is applied to cancer data.

Keywords: Clustering; Ordered variables; Microarray data.

1 Introduction

Finding chromosomal domains of gene expression using microarray technology is a recurrent biological question. From a $n \times p$ microarray dataset (where n is the number of samples and p the number of genes), one can deduce the gene correlation matrix and use it to find chromosomal domains, i.e. small portions of the genome where adjacent genes are highly correlated. These chromosomal domains may be associated to changes in the gene copy number or epigenetic alterations that lead to transcriptional deregulation in cancers. So far, only heuristic algorithms have been proposed, most of them based on sliding window strategies [2,3]. We propose a statistical approach for this problem, where clusters of genes are identified and then tested to detect strongly up-correlated regions.

2 Statistical framework

2.1 Model and objective

We consider a sequence of p ordered variables X^1, \dots, X^p , where $(X^j)^T = (x_1^j, \dots, x_n^j)$ is a vector of n observations. In the microarray experiment context, x_i^j represents the expression level of gene j for sample i . We assume that this sequence breakdowns into K regions (i.e. clusters) $\mathcal{C}_1, \dots, \mathcal{C}_K$, such that

$$\forall j \in \mathcal{C}_k, \quad X^j = A^k + E^j \quad ,$$

where A^k is the cluster information variable, and E^j is a residual variable. We suppose that

$$\forall j \neq j', \forall k \neq k', \quad \text{cov}(A^k, E^j) = 0, \quad \text{cov}(A^k, A^{k'}) = 0, \quad \text{cov}(E^j, E^{j'}) = 0 .$$

In the following, we assume that all variables have been normalized, and we note ρ_k the correlation between two variables of cluster \mathcal{C}_k . Moreover, we assume that for most of the clusters $\rho_k = \rho_0 > 0$, and that for a small subset of clusters $\rho_k > \rho_0$. The goal is then to identify clusters where the correlation is higher than ρ_0 .

To this end, a two-step strategy is proposed. In a first step, we aim at identifying the clusters. In a second step, a test is performed to detect clusters with high correlations.

2.2 Clustering for ordered variable

Assuming variables $X^\ell, \dots, X^{\ell+p_k-1}$ are the p_k variables of cluster \mathcal{C}_k , all these variables are noisy copies of the cluster information variable A^k . Furthermore, information A^k may be estimated with

$$\widehat{A}^k = \frac{1}{p_k} \sum_{j=\ell}^{\ell+p_k-1} X^j = \overline{X}^{(k)}$$

that is (up to a constant) the BLUP of A^k . With these notations, we can define the loss for cluster \mathcal{C}_k by

$$L(\mathcal{C}_k) = p_k - \sum_{j=\ell}^{\ell+p_k-1} \text{cov}(X^j, \overline{X}^{(k)}) .$$

In the ideal case where the noise would be null, all the covariances equal to 1 (since variables are normalized), and the loss is 0. Otherwise, the loss is positive.

We can define the optimal clustering \mathcal{C}^* as the one that satisfies

$$\mathcal{C}^* = \underset{\mathcal{C}}{\text{Argmin}} \sum_{k=1}^K L(\mathcal{C}_k) .$$

With this definition, one only needs to explore all the possible clusterings to find the optimal one. Taking into account that the loss we defined is additive and that clusters only contains adjacent variables, an exhaustive search is possible using dynamic programming (DP). While this provides the optimal clustering, the complexity cost of this algorithm is high ($\mathcal{O}(p^2)$). As an alternative, we propose the use of a Constrained version of the Hierarchical Clustering Algorithm (CHCA), where only adjacent clusters can be merged at each step. The definition of the distance between clusters is directly deduced from the loss function:

$$D(\mathcal{C}_k, \mathcal{C}_\ell) = L(\mathcal{C}_{k\ell}) - L(\mathcal{C}_k) - L(\mathcal{C}_\ell) .$$

Compared with DP, the complexity cost of CHCA is low ($\mathcal{O}(p)$), and the results obtained are very similar (not shown).

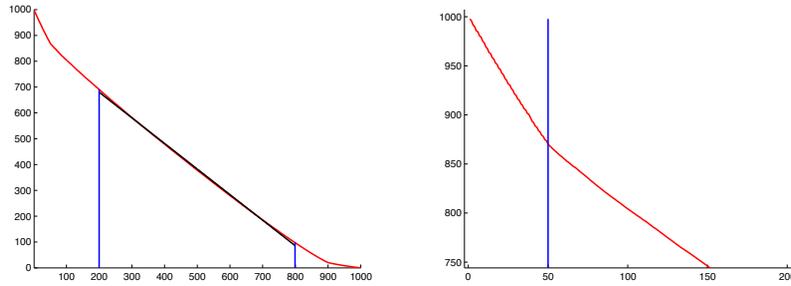


FIGURE 1. **Left:** Clustering curve. The vertical lines represent the bounds for the regression to estimate ρ_0 . **Right:** Zoom on the left part of the curve.

To select the number of clusters, we consider the clustering curve (Figure 1). It represents the history of the clustering, i.e. the loss of the clustering according to the number of clusters. Three phases can be distinguished from the right to the left: first, up-correlated variables are clustered, then ρ_0 -correlation regions are reconstituted, and to finish whole reconstituted clusters are (unwisely) merged together. Each transition is marked by a breakpoint in the slope of the curve, and the breakpoint between phases 2 and 3 indicates the “optimal” number of clusters. To find this last breakpoint, we used the slope break detection method proposed in [1].

2.3 Test for identifying up-correlated regions

For a given cluster, we consider the test $\mathcal{H}_0 : \{\rho_k = \rho_0\}$ vs $\mathcal{H}_1 : \{\rho_k > \rho_0\}$. Assuming A^k and E^j to be gaussian, it is easy to show that

$$n(p_k - L(\mathcal{C}_k)) = \frac{1}{n} \sum_i^n \left(\bar{X}_i^{(k)} - \bar{\bar{X}}^{(k)} \right)^2 \sim \frac{(1 + (p_k - 1)\rho_k)}{np_k} \chi_{(n-1)}^2$$

Thus, the \mathcal{H}_0 hypothesis is rejected for cluster \mathcal{C}_k if $n(p_k - L(\mathcal{C}_k)) > (1 + (p_k - 1)\rho_0)\chi_{n-1, 1-\alpha}^2$. To perform this test, an estimator of parameter ρ_0 is needed. This estimator can be derived using the second phase of the clustering curve. Indeed, it can be shown that the theoretical distance between two groups of variables that belong to the same ρ_0 -correlation cluster is $1 - \rho_0$. Thus, on phase 2 the clustering curve is linear, with slope $1 - \rho_0$. ρ_0 may be estimated by performing a simple regression on phase 2 of the clustering curve (see Figure 1). Since most of the clusters are supposed to have ρ_0 correlations, phase 2 constitutes the major part of the clustering curve. Therefore, we discarded the first and last 20% of the curve, and used the remaining 60% to estimate ρ_0 .

3 Application to bladder cancer data

We consider the chromosome 3 microarray data published in [4]. The dataset consists in microarray measurements for 469 genes in 57 bladder tumors.

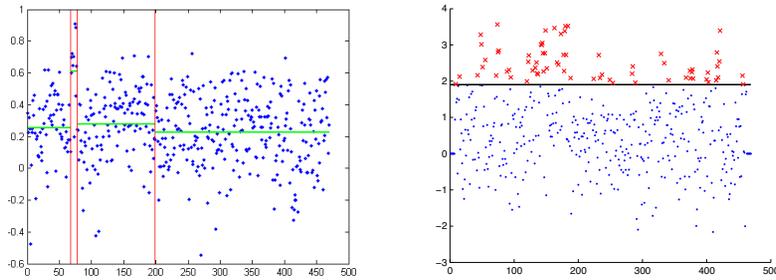


FIGURE 2. **Left:** Result of the clustering strategy. Genes are represented in abscisses, in their chromosomal order. Clusters are delimited with vertical lines. Each point represents the correlation between a variable of cluster k and $\bar{X}^{(k)}$. Horizontal lines in each cluster correspond to the average of the previous correlations. **Right:** Result of the sliding window strategy. Genes whose score is higher than the horizontal line are declared up-correlated with there neighbors.

We applied our strategy to these data and compared it to the sliding window strategy used in the original article. The results are given in Figure 2. The selected number of clusters is 4, and ρ_0 is estimated to 0.06. After testing, only cluster 2 is declared significantly up-correlated. This cluster corresponds to a chromosomal domain that was identified in the original article. As a comparison, sliding window methods provide a gene-by-gene decision that fails to clearly identify the chromosomal domain. The same kind of results may be observed on the other chromosomes (results not shown).

References

- [1] Lavielle, M. (2005). Using penalized contrasts for the change-points problem. *Signal Processing*, **85(8)**, 79-102.
- [2] Reyat, F., *et al.* (2005). Visualizing chromosomes as transcriptome correlation maps: evidence of chromosomal domains containing co-expressed genes—a study of 130 invasive ductal breast carcinomas. *Cancer Research*, **65**, 1376-1383.
- [3] Spellman, P.T., and Rubin, G.M. (2002). Evidence for large domains of similarity expressed genes in the *Drosophila* genome. *Journal of Biology*, **1**, 1-5.
- [4] Stransky, N., *et al.* (2006). Regions of copy number-independent de regulation of transcription in cancer. *Nature genetics*, **38(12)**, 1386-1396.

Empirical likelihood based approach for the inference of the Youden index and associated threshold

Elisa-María Molanes-López¹ and Emilio Letón¹

¹ Department of Statistics, Universidad Carlos III de Madrid, Avda. de la Universidad, 30, 28911 Leganés (Madrid), Spain

Abstract: The Youden index is a widely used measure in the framework of medical diagnostic, where the effectiveness of a biomarker (screening marker or predictor) for classifying a disease status is studied. When the biomarker is continuous, it is important to determine the threshold or cut-off point to be used in practice for the discrimination between diseased and healthy populations. We introduce a new method based on adjusted empirical likelihood for quantiles aimed to estimate the Youden index and its associated threshold. We also include bootstrap based confidence intervals for both of them. In the simulation study, we compare this method with a recent approach based on the delta method. Finally, a real example of prostatic cancer, well known in the literature, is analyzed to provide the reader with a better understanding of the new method.

Keywords: Confidence interval; Empirical likelihood; Optimal cut-off point; ROC curve; Youden index.

Long abstract version: accepted for oral presentation.

Communicating author: Emilio Letón (emilio.leton@uc3m.es)

1 Introduction

Diagnostic tests are often used for classifying diseased and healthy populations. They are based on biomarkers, which can be dichotomous, ordinal or continuous. From here on, we will focus on continuous biomarkers and will assume, without loss of generality, that larger values of the biomarker, denoted by X , are associated with the diseased population.

In this context, a person is classified as ‘diseased’ if the corresponding biomarker value is greater than a given threshold value, and is classified as ‘healthy’ otherwise. Denoting by c that threshold value, there are two important probabilities associated with it: the sensitivity, $q(c)$, and the specificity, $p(c)$. The sensitivity is defined as ‘true positive subjects’, i.e. correctly classified diseased individuals and the specificity as ‘true negative subjects’, i.e. correctly classified healthy individuals. The pairs $(1 - p(c), q(c))$, for all possible threshold values c , are usually drawn in a plot called the

ROC curve. The ROC curve describes graphically the performance of the biomarker under several cut-off points.

A key point in this methodology is to find an optimal threshold, in order to maximize the effectiveness of the biomarker. In most instances, there is an inverse relationship between sensitivity and specificity, in the sense that moving the threshold increases one while decreasing the other. So a kind of balance between sensitivity and specificity is necessary.

There exist two main methods for identifying the optimal threshold: the northwest corner and the Youden index. These two methods can give different cut-off values as Perkins and Schisterman (2006) point out. From here on, we will concentrate on the latter defined by Youden and recently studied by Fluss et al. (2005), Schisterman et al. (2005), Le (2006), and Schisterman and Perkins (2007), among others.

In order to maximize the effectiveness of the biomarker, the Youden index, J , is defined as follows,

$$J = \max\{J(c); c \in \mathfrak{R}\}, \text{ where } J(c) = q(c) + p(c) - 1.$$

The notation X_0 and X_1 will be used to refer to the values of the biomarker on the healthy and diseased populations, respectively. Denoting by F_0 and F_1 their corresponding cumulative distribution functions (cdf's), and by \bar{F}_0 and \bar{F}_1 their complementary ones, it follows that

$$\begin{aligned} J(c) &= \Pr(X_1 > c) + \Pr(X_0 < c) - 1 \\ &= \bar{F}_1(c) + F_0(c) - 1 = \bar{F}_1(c) - \bar{F}_0(c) = F_0(c) - F_1(c). \end{aligned}$$

The manuscript is organized as follows. In Section 2, we propose a new method for computing the optimal J and c , and their confidence intervals. In Section 3, we check their performance under different scenarios and compare it with a recent proposed methodology. In Section 4, the new method is illustrated through a well-known real example.

2 New method

Likelihood-based methods can deal with incomplete data, pool information from different sources and, when there exists extra information from the outside, they can include it as constraints, restricting the domain of the likelihood function, or as prior distributions multiplying the likelihood function. On the other hand, nonparametric estimates avoid the misspecification inherent to parametric model-based estimates. The combination of likelihood and nonparametric methods has been termed in the literature as empirical likelihood (EL). It was first proposed by Thomas and Grunke-meier (1975) and later on, Owen (1990) and other authors have shown the potential of this approach, which nowadays is still an active area of research (see, for instance, Cao and Van Keilegom, 2006, and Molanes-López et al.,

2009). One of the main advantages of empirical likelihood based confidence intervals is that they respect the range of the parameter space, are invariant under transformations and their shape is data-driven.

We present a new approach, based on EL and bootstrapping, for estimating the optimal threshold and the associated Youden index, and their corresponding confidence intervals. However, since knowing the cut-off of the biomarker is more relevant to classify individuals in the medical field, our main focus will be on correctly estimate the optimal threshold. As a byproduct of our method, we get as well an estimate of J .

Before going on more details, we first require to introduce the concept of relative distribution (see Handcock and Morris, 1999, for more details), which is very related to the concept of ROC curve. Specifically, the relative distribution of X_1 with respect to (w.r.t.) X_0 is defined as the cdf of the random variable $Z = F_0(X_1)$, i.e.

$$R_{01}(t) = \Pr(Z \leq t) = \Pr(F_0(X_1) \leq t) = F_1(F_0^{-1}(t)).$$

To provide the reader with some insight on the interpretation of $R_{01}(t)$ for a fixed $t \in (0, 1)$, set $s = R_{01}(t)$, with $s \in (0, 1)$. Then, $F_1(c) = s$ for some c in \mathfrak{R} such that $F_0(c) = t$, i.e. c is the s -th quantile of X_1 and the t -th quantile of X_0 . On the other hand, it is easy to see that $R_{01}(t)$ is a reparametrization of the ROC curve,

$$R_{01}(t) = 1 - ROC(1 - t),$$

where $ROC(t) = \bar{F}_1(\bar{F}_0^{-1}(t))$, for $t \in (0, 1)$, denotes the ROC curve, i.e. the cdf of the random variable $1 - Z$, known in the literature as ‘the placement value’.

Consider $\{X_{0k}\}_{k=1}^{n_0}$ and $\{X_{1k}\}_{k=1}^{n_1}$, two independent samples taken from both populations, X_0 and X_1 , with sample sizes n_0 and n_1 , respectively. Based on these observations, we detail below the 4 steps of our method to obtain estimates of J and c .

Step 1. We obtain $\hat{R}_{01}(t)$, a kernel-type estimate of the relative distribution of X_1 w.r.t. X_0 ,

$$\hat{R}_{01}(t) = \frac{1}{n_1} \sum_{k=1}^{n_1} G\left(\frac{t - F_{0n_0}(X_{1k})}{h_1}\right), \quad (1)$$

and then we find the value t , let say t_0 , that maximizes the distance between $\hat{R}_{01}(t)$ and t . In equation (1) above, $G(x) = \int_{-\infty}^x K(y)dy$, K denotes a kernel function, h_1 is the smoothing parameter, also known as bandwidth, and F_{0n_0} refers to the empirical cdf of X_0 .

Note that, since F_0 is assumed unknown, it is required to estimate it through F_{0n_0} . Consequently, $\hat{R}_{01}(t)$ in (1) can be seen as a traditional kernel-type cdf estimate of Z , based on the pseudosample

$\{F_{0n_0}(X_{1k})\}_{k=1}^{n_1}$ rather than on the unobserved sample $\{F_0(X_{1k})\}_{k=1}^{n_1}$, straightforwardly drawn from Z . Note that $F_{0n_0}(X_{1k})$ above gives the rank of X_{1k} in the healthy sample $\{X_{01}, \dots, X_{0n_0}\}$.

Analogously, interchanging the roles of X_0 and X_1 , we obtain $\hat{R}_{10}(t)$, a kernel-type estimate of the relative distribution of X_0 w.r.t. X_1 , and find the value t , let say t_1 , that maximizes the distance between \hat{R}_{10} and t .

Step 2. With the two values previously computed, t_0 and t_1 , we then apply the adjusted EL method for quantiles proposed by Zhou and Jing (2003), and estimate the t_0 -th quantile of the healthy population, $c_0 = F_0^{-1}(t_0)$, and the t_1 -th quantile of the diseased population, $c_1 = F_1^{-1}(t_1)$.

Specifically, for $i = 0, 1$, we find the value \hat{c}_i , that minimizes the adjusted log-empirical likelihood ratio given by the expression below,

$$\hat{\ell}(c_i) = 2n_i \left(\hat{F}_i(c_i) \log \frac{\hat{F}_i(c_i)}{t_i} + (1 - \hat{F}_i(c_i)) \log \frac{1 - \hat{F}_i(c_i)}{1 - t_i} \right),$$

where \hat{F}_i denotes a kernel-type estimate of F_i ,

$$\hat{F}_i(x) = \frac{1}{n_i} \sum_{k=1}^{n_i} G \left(\frac{x - X_{ik}}{g_i} \right), \quad (2)$$

with g_i the smoothing parameter.

It is interesting to note here that, although most of the empirical likelihood approaches lead to log-likelihood functions implicitly defined by a nonlinear equation, this is not the case for the approach of Zhou and Jing (2003), where a closed form is available for the log-likelihood function.

Step 3. With the two estimates previously computed in Step 2, \hat{c}_1 and \hat{c}_2 , we then propose $\hat{c} = \frac{n_0}{n} \hat{c}_0 + \frac{n_1}{n} \hat{c}_1$ as an estimate of the optimal threshold c , where $n = n_0 + n_1$. Finally, as a byproduct, an estimate of the Youden index is given by $\hat{J} = \hat{F}_0(\hat{c}) - \hat{F}_1(\hat{c})$, where \hat{F}_i has been previously introduced in equation (2).

Step 4. In order to obtain confidence intervals for c and J , we resample independently from both populations and repeat the three steps given above a large number of times, let say B , using the bootstrap resamples. These bootstrap resamples are drawn from smoothed versions of the corresponding empirical cdf's.

Finally, the confidence intervals for the Youden index J and the optimal threshold c are given by the percentile method.

3 Simulation study

A study of interval width and coverage probability is done through a simulation study based on a variety of parametric situations and different sample sizes, including balanced and non-balanced scenarios, similar to those considered by other authors. The results of our new method are compared with the delta method, recently used by Schisterman and Perkins (2007) under parametric assumptions. For the sake of brevity, we will present below the results obtained for the bigamma model.

The bigamma model is given by two gamma distributed random variables, $X_0 \sim \gamma(\alpha_0, \beta_0)$ and $X_1 \sim \gamma(\alpha_1, \beta_1)$, with density functions defined by

$$f_i(\alpha_i, \beta_i, x) = \frac{e^{-x/\beta_i} x^{\alpha_i-1}}{\beta_i^{\alpha_i} \Gamma(\alpha_i)}, \text{ with } x > 0,$$

where $\alpha_i > 0$ and $\beta_i > 0$ are the shape and scale parameters, for $i = 0, 1$, and Γ denotes the gamma function

$$\Gamma(p) = \int_0^{\infty} e^{-x} x^{p-1} dx, \text{ with } p > 0.$$

The specific values, under the bigamma assumption, for the shape and scale parameters of the healthy population were fixed to $\alpha_0 = 1.5$ and $\beta_0 = 1$. However, the parameters of the diseased population were accordingly selected to yield different values of J , as collected in Table 1.

Shape parameter α_1 of X_1	Youden index J			
	$J = 0.4$	$J = 0.6$	$J = 0.8$	$J = 0.9$
$\alpha_1 = 1.5$	2.4828	4.3565	9.7847	19.8020
$\alpha_1 = 2.0$	1.6622	2.7650	5.6517	10.3842

TABLE 1. *Bigamma model: scale parameter of diseased population, β_1 , with $\alpha_0 = 1.5$ and $\beta_0 = 1$.*

The simulations were carried out in MATLAB. For every scenario specified in Table 1, 300 trials were considered. For each trial, a sample of $n_0 = 50$ i.i.d. observations, $\{X_{01}, \dots, X_{0n_0}\}$, and a sample of $n_1 = 50$ i.i.d. observations, $\{X_{11}, \dots, X_{1n_1}\}$, were independently drawn from X_0 and X_1 , respectively. The uniform kernel, K , was considered to estimate the relative distributions involved in the first step of our algorithm, R_{01} and R_{10} , and the cdf's, F_i , for $i = 0, 1$, required to estimate J in Step 3. For these kernel type estimates we considered the following bandwidths, $h_i = n_i^{-1/3}$ and $g_i = 2n_i^{-1/3}$, for $i = 0, 1$, that are of optimal order to estimate cdf's in two-sample and one-sample problems. However, given the regularity conditions required by the adjusted empirical likelihood method for quantiles

of Zhou and Jing (2003), we used in (2) bandwidths given by $n_i^{-1/2}$, for $i = 0, 1$, and a kernel different from the commonly used in nonparametric density estimation,

$$K(x) = \left\{ \frac{21 - 9\sqrt{21}}{8}x^2 + \frac{-3 + 3\sqrt{21}}{8} \right\} 1_{\{|x| \leq 1\}}.$$

From each pair of samples, we generated $B = 299$ bootstrap resamples to obtain 95%-confidence intervals for the optimal cut-off point c and the Youden index J . Although in classical resampling methodology the resamples are drawn from the empirical cdf's, we have used instead kernel type cdf estimates of F_i with gaussian kernel and bandwidths given by $0.2 \max \{ \text{iqr}(\{X_{ik}\}_{k=1}^{n_i}), \text{std}(\{X_{ik}\}_{k=1}^{n_i}) \} n_i^{-1/5}$, for $i = 0, 1$, with iqr and std referring to, respectively, the sample interquartile range and sample standard deviation.

Bigamma model: $CI_{95\%}(c)$ with $(n_0, n_1) = (50, 50)$					
		ELM		Delta method	
α_1	J	coverage(%)	width	coverage(%)	width
1.5	0.4	96.33	1.3734	90.33	1.0598
1.5	0.6	96.00	1.4164	93.00	0.9643
1.5	0.8	99.00	1.3263	93.33	1.3414
1.5	0.9	95.33	1.9039	94.33	1.9033
2.0	0.4	96.33	1.2632	96.33	4.1990
2.0	0.6	98.33	1.2676	96.33	1.1748
2.0	0.8	94.67	1.3755	94.67	1.4047
2.0	0.9	93.33	1.5828	93.33	1.9389

TABLE 2. Coverage probabilities and width averages of $CI_{95\%}(c)$.

We would like to remark here that sometimes, in Step 2 of our algorithm, it may be required to deal with upper and lower quantiles, more extreme than those considered by Zhou and Jing (2003). For instance, for the bigamma model with parameters $\alpha_0 = 1.5$ $\beta_0 = 1$ $\alpha_1 = 2$ and $\beta_1 = 10.38$, which corresponds to a setting of $J = 0.90$, it is necessary to estimate the 0.97-quantile of the healthy population, which can be very challenging, specially if the sample size is small.

We collected in Tables 2 and 3 the results from this simulation study. For the sake of simplicity, we will refer to the new method by ELM (Empirical Likelihood Method). From the results gathered in Tables 2 and 3, we observe that in general the ELM for CI's of both parameters of interest tend to present overcoverage, while those based on the delta method present undercoverage. In terms of width average, the ELM for the CI of c behaves better than the other when the biomarker X separates both populations reasonably well (see, for instance, the results in Table 2 for $J = 0.8, 0.9$).

Bigamma model: $CI_{95\%}(J)$ with $(n_0, n_1) = (50, 50)$					
		ELM		Delta method	
α_1	J	coverage(%)	width	coverage(%)	width
1.5	0.4	95.67	0.3301	95.67	0.2757
1.5	0.6	97.67	0.2860	94.67	0.2432
1.5	0.8	96.67	0.2123	94.33	0.1790
1.5	0.9	96.33	0.1746	93.67	0.1220
2.0	0.4	97.00	0.3286	92.33	0.2829
2.0	0.6	96.67	0.2892	93.00	0.2485
2.0	0.8	94.67	0.2106	93.00	0.1810
2.0	0.9	92.67	0.1675	92.00	0.1219

TABLE 3. Coverage probabilities and width averages of $CI_{95\%}(J)$.

It is also interesting to point out the results collected in the third row of Table 2. While the ELM for estimating c presents the higher observed overcoverage, 99%, it shows a width average shorter than the delta method (a width average of 1.3263 versus 1.3414). On the other hand, an isolated case has been observed for the delta method in the fifth row of Table 2, where an atypical trial had a negative effect on the width average.

However, when estimating J , the width of the ELM for the CI is larger than the width of the delta method. This can be explained due to the fact that our approach is focused on correctly estimating c , and once \hat{c} is computed, \hat{J} is obtained as a byproduct. As it was already observed in the literature, even though J and c are strongly related, a good method for estimating one of them is not necessarily good for the other (see Fluss et al., 2005).

From the results of the simulations, we conclude that the new confidence intervals have good performance in terms of nominal coverage and width, being competitive with the delta method. The delta method is dependent on distributional assumptions, and violations of them can yield substantial bias in estimation. Therefore, we suggest using the new method when the underlying distributions, F_0 and F_1 , are unknown.

4 Example

A real example analyzed in Le (2006) is used to illustrate the application of the new approach. There are 53 patients with prostate cancer: 20 out of them with nodal involvement and 33 without. The biomarker used in this example is the level of acid phosphatase in blood serum ($\times 100$).

It is easy to check that these data do not follow any of the parametric models (binormal or bigamma) studied in Schisterman and Perkins (2007) via the delta method. Consequently, a straightforward application of the delta method would not be possible. First, the appropriate parametric model

should be found, which not always may be possible, and then all the formulation required by the delta method should be rewritten.

After analyzing this example using our nonparametric method, we obtain $\hat{c} = 60.67$ for the optimal threshold and the following confidence interval $CI_{95\%}(c) = (51.40, 67.50)$. The point estimate \hat{c} differs from that obtained in Le (2006), $\hat{c} = 75.00$, who proposed to model the ROC function by parametric Lehmann's alternatives. It is interesting to note that this Lehmann-based estimate is outside our confidence interval. This fact suggests that the assumption of Lehmann's alternatives may not be tenable here.

References

- Cao, R. and Van Keilegom, I. (2006). Empirical likelihood tests for two-sample problems via nonparametric density estimation. *Scandinavian Journal of Statistics*, **34**, 61-77.
- Fluss, R., Faraggi, D. and Reiser, B. (2005). Estimation of the Youden index and its associated cutoff point. *Biometrical Journal*, **47**, 458-472.
- Handcock, M.S. and Morris, M. (1999). *Relative distribution methods in social sciences*. New York: Springer.
- Le, C.T. (2006). A solution for the most basic optimization problem associated with an ROC curve. *Statistical Methods in Medical Research*, **15**, 571-584.
- Molanes-López, E.M., Van Keilegom, I. and Veraverbeke, N. (2009). Empirical likelihood for non-smooth criterion functions. *Scandinavian Journal of Statistics* (in press).
- Owen, A.B. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, **18**, 90-120.
- Perkins, N.J. and Schisterman, E.F. (2006). The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American Journal of Epidemiology*, **163**, 670-675.
- Schisterman, E.F. and Perkins, N.J. (2007). Confidence intervals for the Youden index and corresponding optimal cut-point. *Communications in Statistics - Simulation and Computation*, **36**, 549-563.
- Thomas, D.R. and Grunkemeier, G.L. (1975). Confidence interval estimation of survival probabilities for censored data. *Journal of the American Statistical Association*, **70**, 865-871.
- Zhou, W. and Jing, B.Y. (2003). Adjusted empirical likelihood method for quantiles. *Annals of the Institute of Statistical Mathematics*, **55**, 689-703.

A Beta-Poisson Model for Underreporting

Gerhard Neubauer¹ and Gordana Djuraš¹

¹ Institute of Applied Statistics, JOANNEUM RESEARCH, Graz, Austria

Abstract: Underreporting in register systems can be analyzed applying a binomial approach, where observed counts are used to estimate both parameters. The usual moment and likelihood methods fail when overdispersion is present. Neubauer & Friedl (2006) introduced a Beta-Binomial model, which results from assuming $P \sim \text{Beta}(\gamma, \delta)$ for a random reporting probability. In this paper we propose to consider both binomial parameters as random, by using $L \sim \text{Poisson}(\lambda)$ in addition to Neubauer & Friedl (2006). The resulting marginal is a Beta-Poisson distribution with parameters λ, γ and δ . The expected number of cases is λ and the expected reporting probability is $\gamma/(\gamma + \delta)$. We introduce a regression approach and give a approximate maximum likelihood method for parameter estimation. The properties of the method are investigated in a simulation study and finally crime data are analyzed.

Keywords: Regression, Laplace approximation, Crime data

1 Introduction

Underreporting is a problem in data collection, when events are counted and for some reason errors occur. As a consequence the mean of the observed counts is smaller than the true mean λ . Using a Binomial model the mean of the observed counts is $\mu = \lambda\pi$, with π the reporting probability, and both parameters to be estimated. Neubauer & Friedl (2006) introduced a regression approach for the Binomial model and - to adopt for overdispersion - also for a Beta-Binomial model. The Binomial and a Beta-Binomial regression model are suited for a wide range of applications. However, if the sample variance is larger than the sample mean the binomial approach fails to give reasonable estimates. For this kind of data Neubauer & Djuraš (2008) proposed a regression model based on the Generalized Poisson distribution (Consul, 1989). This model allows to handle Poisson under- and overdispersion, as it covers the binomial, Poisson and the negative binomial case. In this paper we propose a further extension of the binomial approach leading to a Beta-Poisson regression model. The Beta-Poisson distribution is also known as "Type H_1 " distribution (Johnson, Kemp & Kotz, 2005), and is usually treated in the context of contagious distributions. Inference is based upon maximum likelihood estimation, where the involved confluent hypergeometric function is approximated by Laplace technique. The

performance of this model is investigated by a simulation study and in an application to real data.

2 Regression Models

Let y_t be a sample of counts ($t = 1, \dots, T$), which are the reported cases of some register system. Further let λ denote the total number of cases and π the reporting probability, then $E(Y_t) = \mu = \lambda\pi$ is the mean model. For the $Y_t \sim \text{Binomial}(\lambda, \pi)$ we have the Binomial model for the estimation of the total number λ with the mean-variance relation $\text{var}(Y_t) = \mu - \mu^2/\lambda$. Allowing for larger variability is possible by treating parameters as random variables. The counts now have a conditional binomial distribution. For $Y_t|P \sim \text{Binomial}(\lambda, P)$ and $P \sim \text{Beta}(\gamma, \delta)$ we obtain the well-known beta-binomial as marginal distribution of Y_t , with $\text{var}(Y_t) = (\mu - \mu^2/\lambda)\phi$, and $\phi = (\lambda + \gamma + \delta)/(1 + \gamma + \delta) \geq 1$. For $Y_t|L \sim \text{Binomial}(L, \pi)$ together with $L \sim \text{Poisson}(\lambda)$ we have $Y_t \sim \text{Poisson}(\lambda\pi)$. For cases with even more variation we take the conditional approach

$$\begin{aligned} Y_t|P &\sim \text{Poisson}(\lambda P) && \text{and} \\ P &\sim \text{Beta}(\gamma, \delta) && \text{resulting in} \\ Y_t &\sim \text{Beta-Poisson}(\lambda, \gamma, \delta). \end{aligned} \tag{1}$$

The same result is obtained for the conditional model

$$\begin{aligned} Y_t|L &\sim \text{Beta-Binomial}(L, \gamma, \delta) && \text{and} \\ L &\sim \text{Poisson}(\lambda). \end{aligned}$$

The Beta-Poisson has the first two moments $E(Y_t) = \mu = \lambda\pi$ and $\text{var}(Y_t) = \mu\phi$ with $\pi = \gamma/(\gamma + \delta)$ and $\phi = 1 + \lambda(1 - \pi)/(1 + \gamma + \delta)$.

More flexible models are at hand if $E(Y_t) = \mu_t$ is allowed and $\lambda_t = \exp(x'_t\beta)$ is used to make parameters identifiable. x_t is a d -vector of known regressors and β is the corresponding vector of unknown parameters. The likelihood contribution of the t -th observation is now

$$L(\beta, \gamma, \delta|y_t, x_t) = \frac{\lambda_t^{y_t} B(a_t, \delta)}{y_t! B(\gamma, \delta)} {}_1F_1[a_t; c_t; -\lambda_t], \tag{2}$$

where $B(\cdot)$ is the beta function, ${}_1F_1[\cdot]$ denotes the confluent hypergeometric function, $a_t = y_t + \gamma$ and $c_t = y_t + \gamma + \delta$.

3 Approximation of the Beta-Poisson Likelihood

The confluent hypergeometric function ${}_1F_1$ has two representations that are both unsuited for the analytical treatment needed in maximum likelihood estimation. Using the integral representation

$${}_1F_1[a_t; c_t; -\lambda_t] = \frac{1}{B(a_t, \delta)} \int_0^1 p^{a_t-1} (1-p)^{\delta-1} e^{-\lambda_t p} dp = B(a_t, \delta)^{-1} I_t \tag{3}$$

equation (2) becomes $L(\beta, \gamma, \delta|y_t, x_t) = \lambda_t^{y_t} / (y_t! B(\gamma, \delta)) I_t$. For (3) Butler & Wood (2002) derive a Laplace approximation. They investigate the accuracy of the approximation in simulations and find that it is "extremely high". Using their results the raw approximation is given as

$${}_1\widehat{F}_1[a_t; c_t; -\lambda_t] = \frac{1}{B(a_t, \delta)} \frac{(2\pi)^{1/2} p_s^{a_t} (1 - p_s)^\delta}{[a_t(1 - p_s)^2 + \delta p_s^2]^{1/2}} \exp(-\lambda_t p_s), \tag{4}$$

where $p_s = p_s(\lambda_t) = 2a_t / (c_t + \lambda_t + \sqrt{(c_t + \lambda_t)^2 - 4\lambda_t a_t})$. The calibrated approximation is defined as

$${}_1\widetilde{F}_1[a_t; c_t; -\lambda_t] = \frac{{}_1\widehat{F}_1[a_t; c_t; -\lambda_t]}{{}_1\widehat{F}_1[a_t; c_t; 0]},$$

where ${}_1\widehat{F}_1[a_t; c_t; 0]$ uses $p_0 = p_s(\lambda_t = 0) = a_t / c_t$ instead of p_s in (4). Apparently ${}_1\widetilde{F}_1[\cdot] = \widetilde{I}$. The approximated likelihood uses ${}_1\widetilde{F}_1$ instead of ${}_1F_1$. Estimation and inference is based on the approximated log-likelihood of the beta-Poisson regression model. For the beta parameters we use parameterization $\alpha = \text{logit}(\pi)$ and $\theta = \gamma + \delta$. The estimation algorithm cycles between MLE of α and β given θ , and the method of moments estimation of θ given α and β . The parameter θ is related to the Poisson overdispersion though $\phi = 1 + \lambda(1 - \pi) / (1 + \theta)$, and hence we estimate it by the method of moments.

4 A simulation study

To investigate the behavior of the beta-Poisson regression model we performed a simulation study. In all situations we used a model with a trend and a seasonality component, reflecting a typical situation with crime data. The data are simulated from the beta-Poisson distribution using $\lambda_t = \exp(\beta_0 + T_t + S_t)$, where $T_t = \beta_1 t + \beta_2 t^2 + \beta_3 \sin(\pi t / 2\psi)$ is the trend function, $S_t = \beta_4 \cos(2\pi t / \psi)$ the seasonality, $\beta = (4, 0.01, -0.00005, 0.2, -0.1)$ and $\psi = 365.25 / 7$ tunes the trigonometric functions. For $t = 1, \dots, 209$ we have about four seasons in the simulated data. Figure 1 gives an example for simulated data and the model for μ (solid) and λ (dashed).

As we are not interested in modelling reporting systems that miss more than half of the cases we focus on situations with $\pi \geq 0.5$. The parameters of the beta distribution were varied to obtain different data types. For large values of γ and δ we have $\text{var}(P) = \sigma^2 = \pi(1 - \pi) / (1 + \theta) \rightarrow 0$ and the beta-Poisson distribution approaches the Poisson limit. The same holds for $\pi \rightarrow 0$, when $E(Y) = \text{const}$. The Poisson model $E(Y_t) = \lambda_t \pi$ is not identifiable and estimation results in $\hat{\lambda}_t \rightarrow \infty$ with $\hat{\pi} \rightarrow 0$ in the limit. From the simulation we expect to get some idea when this situation occurs in finite samples. Moreover we want to find out if this behavior holds equally

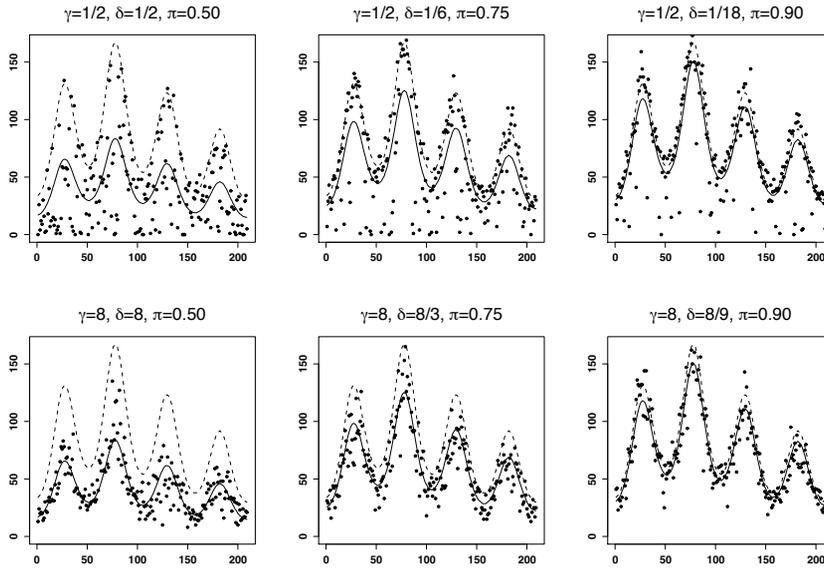


FIGURE 1. Examples of simulated data with true trends for the mean (solid) and the total number of cases (dashed)

TABLE 1. Estimated reporting probabilities from the simulation study

	(γ, δ)	π	σ^2	$\hat{\pi}$	$\hat{\sigma}^2$	CI(π)	ϕ
a	(1/2, 1/2)	0.50	0.125	0.541	0.134	0.484; 0.596	20.291
b	(1, 1)	0.50	0.083	0.528	0.089	0.475; 0.580	13.860
c	(4, 4)	0.50	0.028	0.532	0.032	0.474; 0.589	5.287
d	(8, 8)	0.50	0.015	0.566	0.019	0.498; 0.632	3.269
e	(1/2, 1/6)	0.75	0.113	0.769	0.120	0.713; 0.817	12.574
f	(1, 1/3)	0.75	0.080	0.745	0.084	0.692; 0.792	9.267
g	(4, 4/3)	0.75	0.030	0.728	0.029	0.672; 0.778	4.046
h	(8, 8/3)	0.75	0.016	0.710	0.014	0.641; 0.771	2.653
i	(1/2, 1/18)	0.90	0.058	0.870	0.058	0.825; 0.905	5.960
j	(1, 1/9)	0.90	0.043	0.872	0.042	0.829; 0.906	4.655
k	(4, 4/9)	0.90	0.017	0.846	0.015	0.794; 0.887	2.417
l	(8, 8/9)	0.90	0.009	0.828	0.008	0.753; 0.883	1.780

for different values of π . Setting $\pi = (0.50, 0.75, 0.90)$ and $\gamma = (1/2, 1, 4, 8)$ we have twelve data situations, with the beta variability σ^2 in the range of $(0.009, 0.125)$. For each of the twelve settings (a-l) $R = 100$ samples were drawn.

Table 1 gives the simulation settings for the beta parameters and estimated moments of the beta component. Replacing the parameters with their es-

estimates in π and σ^2 gives the estimates in Table 1. The confidence interval is obtained using the delta method.

The regression parameters are well estimated in all cases and therefore details are omitted for brevity. We find that the reporting probability can be well estimated for $\pi \geq 0.5$. The confidence interval $CI(\pi)$ covers the true value in most cases. Only for situations where $\sigma \rightarrow 0$ we observe non-coverage (k,l in Table 1).

5 Application to real data

Neubauer & Djuraš (2008) used data from the Austrian online crime register SIMO to give an example for the performance of the Generalized Poisson (GP) model. For shop lifting in an Austrian region and bicycle theft in a city the estimates of the reporting probability were $\hat{\pi} = 0.69$ and $\hat{\pi} = 0.67$. Applying the Beta-Poisson model we obtain $\hat{\pi} = 0.65$ and $\hat{\pi} = 0.52$. The result from the two models for the bicycle data are different and there is some evidence that the Beta-Poisson model is more adequate in this case. Figure 2 shows the results for the Beta-Poisson model.

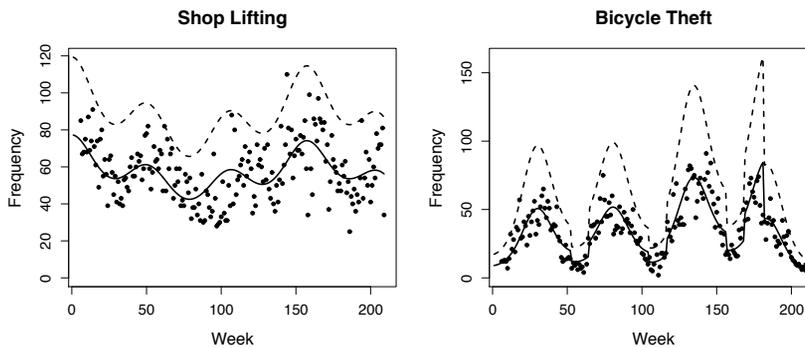


FIGURE 2. Two examples of Austrian crime data with estimated trends for the mean (solid) and the total number of cases (dashed)

6 Summary

A Beta-Poisson model is developed to allow for the estimation of binomial parameters, when data show large overdispersion. The estimation relies on an approximate likelihood that results from replacing a confluent hypergeometric function by its Laplace approximation. The proposed model and estimation technique show good performance in a simulation study where the expected reporting probability π is larger than 0.5. Finally the model

is applied to two examples of Austrian crime data and the results are compared to those obtained in Neubauer & Djuraš (2008). For one example the results differ and therefore an inferential procedure to decide between models for underreporting is needed. Usual likelihood approaches are not applicable but non-nested testing techniques seem promising and are subject of future research.

Acknowledgements

This work was generously funded by the Austrian Ministry for Transport, Innovation and Technology within the program "Zielvereinbarung 2007-2008".

References

- Butler, R.W. and Wood, A.T.A. (2002). Laplace Approximations for Hypergeometric Functions with Matrix Argument, *The Annals of Statistics*, 30, 1155-1177.
- Johnson, N.L., Kemp, A.W. and Kotz, S. (2005). *Univariate Discrete Distributions*, Hoboken: Wiley.
- Neubauer, G. and Djuraš, G. (2008). A Generalized Poisson Model for Underreporting. In: *Proceedings of the 23rd International Workshop on Statistical Modelling*, 7-11 July, 2008, Utrecht, Netherlands.
- Neubauer, G. and Friedl, H. (2006). Modelling sample sizes of frequencies. In: *Proceedings of the 21st International Workshop on Statistical Modelling*, 3-7 July 2006, Galway, Ireland.

Bayesian Estimation of a Beta-Poisson Model

Gerhard Neubauer¹, Michaela Dvorzak¹ and Helga Wagner²

¹ Institute of Applied Statistics, JOANNEUM RESEARCH, Graz, Austria

² Department of Applied Statistics, Johannes Kepler University, Linz, Austria

Abstract: The size of underreporting in register systems can be estimated by using a binomial model, where both parameters λ and π have to be estimated. Both moment and likelihood methods give negative estimates, when overdispersion is present. Therefore binomial mixtures and regression have been proposed to accommodate for overdispersion. In this paper we develop a Bayesian approach for the estimation of a beta-Poisson model where λ is the true number of events subject to underreporting and π the expected reporting probability.

Keywords: Regression, auxiliary mixture sampling, underreporting, crime data.

1 Introduction

The counting of events may be subject to systematic biases resulting in under- or overreporting. One prominent example is the reported number of crimes, which is in general considered as being lower than the number of committed crimes. For such cases μ , the mean of the observed counts, is smaller than the true mean λ . Using a binomial model we can express this as $\mu = \lambda\pi$, where π is the reporting probability. We denote this model as $Y \sim \text{Binomial}(\lambda, \pi)$, where both parameters are to be estimated.

Real data often exhibit more variation than the binomial model can handle. Therefore mixture models have been proposed by Neubauer & Friedl (2006) and Neubauer & Djuraš (2008), where either π or λ was assumed to be random. Neubauer & Djuraš (2009) propose a model where both parameters are considered as random and

$$Y|L, P \sim \text{Binomial}(L, P).$$

Assuming $L \sim \text{Poisson}(\lambda)$ together with $P \sim \text{Beta}(\gamma, \delta)$ we obtain the beta-Poisson distribution as marginal distribution for the observed count y , where

$$E(Y) = \mu = \lambda \frac{\gamma}{\gamma + \delta} = \lambda\pi$$

and

$$\text{var}(Y) = \mu \left(1 + \frac{\lambda(1 - \pi)}{1 + \gamma + \delta} \right) = \mu\phi.$$

Hence, we have the same marginal mean as with the binomial distribution, but $\text{var}(Y) > \mu$, while the binomial is restricted to $\text{var}(Y) < \mu$.

For regression we use $E(Y_t) = \lambda_t \pi$, $t = 1, \dots, T$ and $\log(\lambda_t) = x_t' \beta$, where x_t is a d -dimensional vector of regressors and β the corresponding vector of parameters (see Neubauer & Djuraš (2009) for further details).

2 Bayesian inference

As the posterior of the parameter vector $\vartheta = (\beta, \gamma, \delta)$ in the beta-Poisson regression model is intractable, direct Bayesian inference is not possible. We therefore use an MCMC algorithm with data augmentation. Conditional on the reporting probabilities P_t ,

$$Y_t | P_t \sim \text{Poisson}(\lambda_t P_t),$$

i.e. we deal with a Poisson regression model with offset P_t where we can apply the improved auxiliary mixture sampler proposed in Frühwirth-Schnatter et al. (2008) to estimate the vector of regression coefficients. The parameters of the beta distribution γ and δ are estimated using the reparameterization $\pi = \gamma/(\gamma + \delta)$ and $\theta = \gamma + \delta$ by two Metropolis Hastings steps. The auxiliary mixture sampler for Poisson counts is built upon the introduction of two sequences of missing data. This leads to a normal posterior distribution for β that, once we conditioned on the augmented data, enables straightforward Gibbs sampling. In a first step inter-arrival times $\tau_t = (\tau_{t1}^*, \tau_{t2}^*)$ of an assumed Poisson process in the time interval $[0, 1]$ with intensity $\lambda_t P_t$ are introduced. As the arrival times of the underlying Poisson process follow an exponential distribution, we have $\tau_{t1}^* = \xi_{t1}/\lambda_t P_t$ with $\xi_{t1} \sim \text{Exponential}(1)$ and $\tau_{t2}^* = \xi_{t2}/\lambda_t P_t$ with $\xi_{t2} \sim \text{Gamma}(y_t, 1)$ where τ_{t1}^* is the waiting time between the last jump before and the first jump after 1 and τ_{t2}^* the arrival time of the last jump before 1. Reformulating these equations yields

$$-\log \tau_{tj}^* - \log P_t = \log \lambda_t + \varepsilon_{tj} = x_t' \beta + \varepsilon_{tj},$$

for $j = 1, 2$, where $\varepsilon_{t1} = -\log \xi_{t1}$ and $\varepsilon_{t2} = -\log \xi_{t2}$.

The non-normal error terms ε_{t1} and ε_{t2} are approximated by Gaussian mixtures and the data are augmented by the latent indicators of the mixture component $r_t = (r_{t1}, r_{t2})$. Hence, conditioning on both sequences $\tau = \{\tau_1, \dots, \tau_T\}$ and $r = \{r_1, \dots, r_T\}$ leads to a Gaussian regression model. Based on this regression model, β is sampled from a multivariate normal distribution, for details see Frühwirth-Schnatter et al. (2008).

However, Gibbs sampling is not feasible for the parameters π and θ of the beta distribution, as their posterior distributions are not of closed form. Sampling of π and θ requires two Metropolis Hastings steps, where we use a uniform random walk for π and a log-normal random walk for θ .

Sampling the reporting probabilities P_t is straightforward, conditional on knowing the dark figure D_t as

$$P_t | (y_t, D_t) \sim \text{Beta}(\gamma + Y_t, \delta + D_t).$$

Hence, also the dark figures $D_t = L_t - Y_t$, $t = 1, \dots, T$ have to be sampled from Poisson $(\lambda_t(1 - P_t))$.

The sampling scheme of our MCMC algorithm therefore involves the following steps:

1. Draw the vector of regression coefficients β and the dark figures D_t , $t = 1, \dots, T$,
2. draw π and θ ,
3. draw the reporting probabilities P_t , $t = 1, \dots, T$.

3 Application to simulated data

The performance of the MCMC method is analyzed in a simulation study, where the degree of Poisson overdispersion is varied by the choice of the beta parameters. We used 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 $t = 2, 4, 6, \dots, 200$, $\beta = (4, 0.01, -0.00005, 0.1)$ and $\psi = 365.25/7$, which tunes the trigonometric function. To obtain different data types the parameters of the beta distribution (γ, δ) and π were varied. Setting $\gamma = (1/2, 4, 16)$ and $\pi = (0.1, 0.3, 0.5, 0.7, 0.9)$ we have 15 data situations.

We use a multivariate normal prior for the regressors with zero mean and covariance matrix $\Sigma = \sigma^2 I$ with $\sigma^2 = 100$ and an uninformative Beta(1,1) prior for π to express vague prior information. For θ we found it necessary to use slightly informative priors depending on the magnitude of the true values, e.g. for the true value of θ in the range $(0, 15)$ we used Gamma(2, 0.2) with expectation 10 and variance 50. For each setting $S = 100$ samples were drawn and per sample the sampler was run 20000 times with a burn-in of 10000 runs.

The overall finding is that the sampler estimates the slopes of the regression parameters very well and here convergence is observed in all different settings. Also the magnitudes of the intercept β_0 and the reporting probability π are captured well. The best results are obtained for large beta variability, i.e. $\gamma = 1/2$. The acceptance rates of the Metropolis Hastings steps are sensible in all settings. For $\pi = (0.5, 0.7, 0.9)$ the traceplots and the autocorrelation functions (ACF) of the draws indicate fast convergence, whereas for $\pi = (0.1, 0.3)$ we observe convergence problems especially for $\pi = 0.1$. The chain does not converge for settings with small beta variability, i.e. $\gamma = 4$ and $\gamma = 16$ and thus the properties of the estimates remain unclear.

Figure 1 shows the traceplots and the ACF for the draws of the parameters in two situations: Fast convergence on the left-hand side, where $\gamma = 1/2$ and $\pi = 0.7$, and slow convergence on the right-hand side, where $\gamma = 4$ and $\pi = 0.7$. The burn-in is left of the dashed vertical lines in the traceplots.

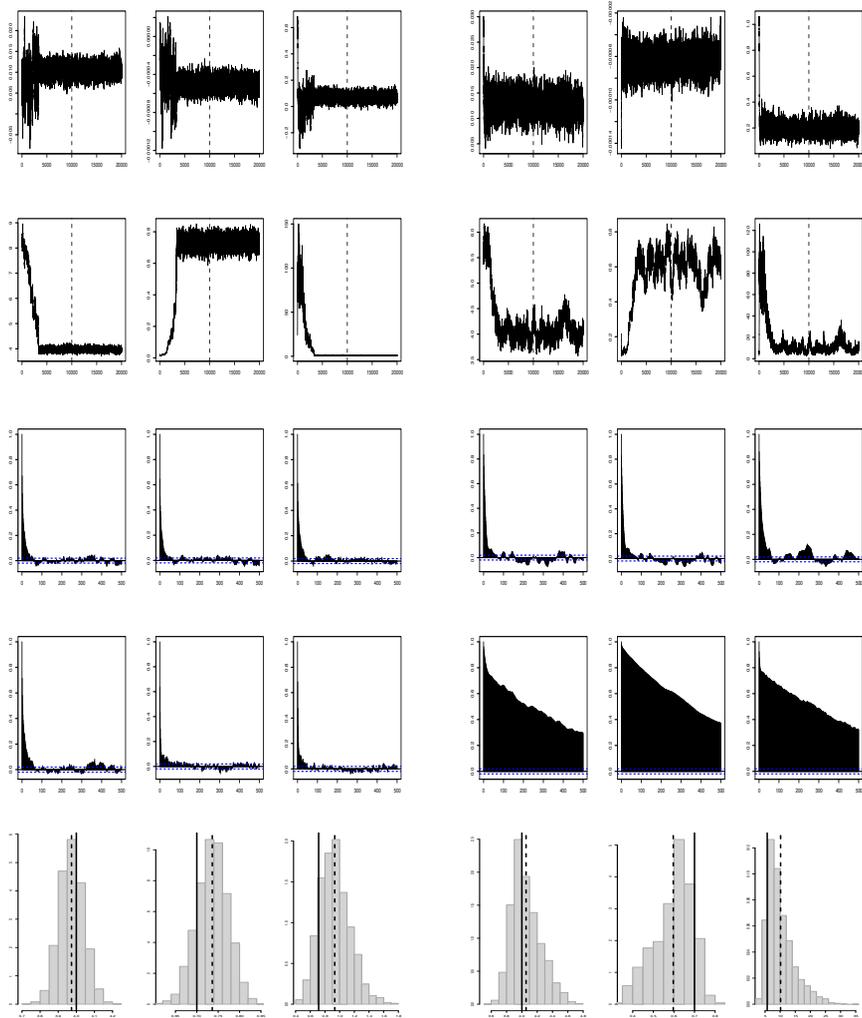


FIGURE 1. Traceplots, ACF and histograms for two examples: fast convergence on the left-hand side ($\gamma = 1/2$, $\pi = 0.7$), and slow convergence on the right-hand side ($\gamma = 4$, $\pi = 0.7$). Row 1: traceplots of slope parameters; row 2: traceplots of β_0 , π , and θ ; row 3: ACF of slope parameters; row 4: ACF of β_0 , π , and θ ; row 5: histograms of β_0 , π , and θ (true value - solid line, estimate - dashed line).

The bottom of Figure 1 gives the histograms of the draws after burn-in for

β_0 , π , and θ . The solid vertical line indicates the location of the true value and the dashed vertical shows the position of the estimate.

To get more information on the convergence of the method the sampler was rerun for all settings of $\gamma = 4$ with 100000 iterations and a burn-in of 20000 runs. This did not substantially change our first impression: slow convergence for β_0 , π and θ , while their magnitudes are captured well. Hence, we conclude that the sampler has some power to find proper estimates, but its performance becomes poor if the Poisson overdispersion parameter ϕ is small or $\pi \rightarrow 0$.

4 Application to real data

The results from the simulation experiment show that the sampler has some power to find sensible estimates if the data are from the beta-Poisson distribution. Here we want to study the behavior of the sampler for real data from the Austrian crime register SIMO. For comparison we have results from Neubauer & Djuraš (2009) who applied MLE to this data. For shop lifting in an Austrian region and bicycle theft in a city the estimated reporting probabilities were $\hat{\pi} = 0.65$ and $\hat{\pi} = 0.52$. We applied the sampler to both data sets with 100000 iterations and the estimates were obtained after removing a burn-in of 25000. The acceptance rates of the Metropolis Hastings steps are sensible for the use of random walk proposals. Traceplots and ACF of the draws indicate convergence problems for β_0 , π and θ . The estimated reporting probabilities are $\hat{\pi} = 0.58$ for the shop lifting and $\hat{\pi} = 0.44$ for the bicycle theft data. Thus they are both lower than the ML estimates. As the sampling variability of π is very large in both cases, we think that the observed difference is not substantial.

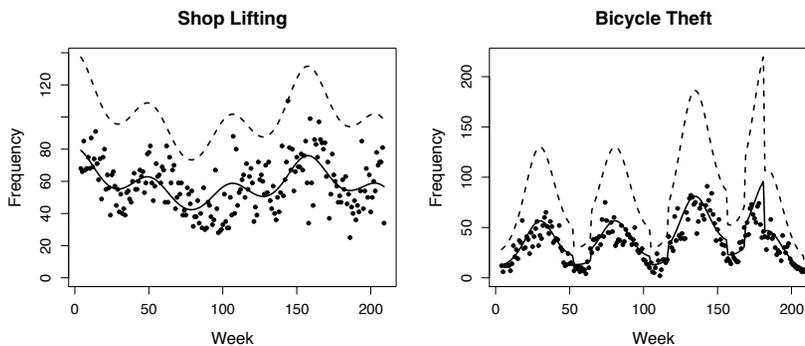


FIGURE 2. Two examples of Austrian crime data with estimated functions for the mean μ (solid) and the total number of events λ (dashed).

Thus, the application to real data yields similar findings as to the simulation study: sensible point estimates, but slow convergence for β_0 , π and θ , and

uncertain inference. Figure 2 shows the estimated functions for the two data sets.

5 Conclusion

A Bayesian estimation procedure for the beta-Poisson regression model to account for overdispersed data is proposed. The estimation technique consists of a Gibbs sampling step for the regressors that results from applying the improved auxiliary mixture sampler and Metropolis Hastings steps for the beta-related quantities. The behavior of the developed MCMC method is investigated in a simulation study that shows good performance for data situations with large beta variability. However, for settings with small beta variability we observe convergence problems, although the magnitude of the parameters is captured well. The method is also applied to examples of Austrian crime data to compare the results to those obtained by MLE in Neubauer & Djuraš (2009). As for the simulated data, we get poor convergence diagnostics for the parameters β_0 , π and θ but their estimated size is sensible. Hence, alternative sampling approaches to handle convergence problems and to improve the performance are needed and are subject of our future work.

Acknowledgments: We would like to thank Sylvia Frühwirth-Schnatter for providing the Matlab routines of the improved auxiliary mixture sampler.

This work was generously funded by the Austrian Ministry for Transport, Innovation and Technology within the program "Zielvereinbarung 2007-2008".

References

- Frühwirth-Schnatter, S., Frühwirth, R., Held, L. and Rue, H. (2008). Improved Auxiliary Mixture Sampling for Hierarchical Models of Non-Gaussian Data. *IFAS Research Paper Series 2008-34*. JKU, Linz, Austria.
- Neubauer, G. and Djuraš, G. (2008). A Generalized Poisson Model for Underreporting. In: *Proceedings of the 23rd International Workshop on Statistical Modelling*, 7-11 July, 2008, Utrecht, Netherlands.
- Neubauer, G. and Djuraš, G. (2009). A Beta-Poisson Model for Underreporting. *Proceedings of the 24th International Workshop on Statistical Modelling*, 20-24 July, 2009, Ithaca, USA.
- Neubauer, G. and Friedl, H. (2006). Modelling sample sizes of frequencies. In: *Proceedings of the 21st International Workshop on Statistical Modelling*, 3-7 July, 2006, Galway, Ireland.

On a Family of Distributions in the Context of Quantile Regression

Angela Noufaily¹ and Chris Jones²

¹ Department of Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK. Email: a.noufaily@open.ac.uk.

² Department of Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK. Email: m.c.jones@open.ac.uk.

Abstract: An iterative method for maximum likelihood estimation of the parameters of a generalized gamma distribution is presented; all parameters are unknown and none of them fixed. The work is extended to quantile regression. Comparison with other optimization methods offered by software such as R is done using simulations. A new 3-parameter (2 of them being shape parameters) distribution is proposed; one of the shape parameters controls the part of the distribution close to zero and the other controls the tail of the model.

Keywords: Generalized Gamma Distribution; Maximum Likelihood Estimation.

1 Introduction

Given an observed set of regression data (X_i, Y_i) ($i = 1, \dots, n$), we would like to answer questions such as: How does Y behave with respect to X ? What best shape or model can we attribute to the data, and what limiting curves define its percentiles? In regression analysis the dependent variable Y in the regression equation is modeled as a function of the independent variables X , corresponding parameters, and a random variable error term. The error term represents unexplained variation in the dependent variable. Most commonly this error term is considered as normally distributed. However, what if the unexplained variation in the dependent variable is not symmetric? From here comes the search for distributions with parameters that control skewness. The 3-parameter generalized gamma is our proposed distribution, applicable to response data that take positive values only. In the next sections, we present an iterative procedure for its maximum likelihood estimation and an extension to quantile regression. As Koenker (2005) mentions: “Quantile regression is intended to offer a comprehensive strategy for completing the regression picture.”. Inspired by the generalized gamma, we propose a new 3-parameter distribution and display some of its properties.

2 The Generalized Gamma Distribution

The generalized gamma is a 3-parameter univariate unimodal continuous life distribution that takes as special cases some of the known life distributions such as the exponential, Weibull, and gamma distributions. Being a rich family of distributions, it was recently a subject of interest to researchers. Lately, Cox *et al* (2007) used it to study survival after a diagnosis of clinical AIDS during different eras of HIV therapy.

2.1 The Probability Density Function and Distribution Function

The generalized gamma probability density function, as first defined by Stacy (1962), is

$$f(t) = \frac{\beta}{\theta^\alpha \Gamma(\frac{\alpha}{\beta})} t^{\alpha-1} e^{-(\frac{t}{\theta})^\beta}, \quad (1)$$

where θ is a scale parameter and α and β are shape parameters. θ , β , α , and $t > 0$. Different transformations and reparametrizations were applied to the pdf form (1) through history, hence obtaining different versions of the density function. For example, by setting $k = \frac{\alpha}{\beta}$, we obtain another form of the density function that is more friendly to work with. The cumulative distribution function in this case is

$$F(t) = \frac{\Gamma_{(\frac{t}{\theta})^\beta}(k)}{\Gamma(k)},$$

known as the regularized gamma function. The numerator is the incomplete gamma function defined in Abramowitz and Stegun (1965) as $\Gamma_a(k) = \int_0^a e^{-z} z^{k-1} dz$, and the denominator is the well-known gamma function.

2.2 Maximum Likelihood Parameter Estimation

Prentice (1974) parametrized the generalized gamma distribution in a way that enabled him to perform maximum likelihood estimation efficiently. Lawless (1980) reparametrized the model in a similar, but slightly different way. Both mentioned approaches assume one of the parameters is known and estimate the others on this basis. Following Lawless (1980), but assuming all parameters are unknown, we develop an iterative procedure for maximum likelihood estimation of a generalized gamma distribution.

Let us look at the maximum likelihood parameter estimation problem for the generalized gamma distribution approached through the distribution of $Y_i = \log T_i$ parametrised as

$$f(y) = \frac{k^{k-\frac{1}{2}}}{\sigma\Gamma(k)} \exp\left\{\sqrt{k}w - ke^{w/\sqrt{k}}\right\}, \tag{2}$$

where $w = \frac{y-\mu}{\sigma}$. Note that $-\infty < \mu < \infty$ and $\sigma, k > 0$.

The log likelihood is

$$n \left\{ -\log(\sigma) + \left(k - \frac{1}{2}\right) \log(k) - \log \Gamma(k) + \sqrt{k} \frac{\bar{Y} - \mu}{\sigma} - \frac{k}{n} \exp\left(\frac{-\mu}{\sigma\sqrt{k}}\right) \sum_{i=1}^n \exp\left(\frac{Y_i}{\sigma\sqrt{k}}\right) \right\}.$$

The pdf form (2) is a “location-scale” version of the density function of the logarithm of a generalized gamma random variable, whereby μ and σ are the location and scale parameters respectively, and k is the shape parameter. Starting by the usual way of differentiating the log likelihood with respect to the three parameters in turn, we obtain 3 score equations. We set the score equations equal to zero and simplify them to obtain relations that have to be solved for finding parameter estimates \hat{k} , $\hat{\sigma}$ and $\hat{\mu}$ for k , σ and μ respectively.

Let

$$S_j \equiv \frac{1}{n} \sum_{i=1}^n Y_i^j \exp\left(\frac{Y_i}{\sqrt{k}\sigma}\right); \quad j = 0, 1, 2.$$

The 3 simplified relations are

$$\exp(\mu) = S_0^{\sigma\sqrt{k}}, \tag{3}$$

$$R(\sigma) \equiv \frac{S_1}{S_0} - \bar{Y} - \frac{\sigma}{\sqrt{k}} = 0 \tag{4}$$

and

$$T_L(k) \equiv \log(k) - \psi(k) - \frac{L}{\sqrt{k}}, \tag{5}$$

where $L = (\mu - \bar{Y})/\sigma > 0$ when μ satisfies (3), as can be shown via Jensen’s inequality.

Each of the equations (3), (4), and (5) proves to have one single root $\hat{\mu}$, $\hat{\sigma}$ and \hat{k} respectively. Therefore, we solve them iteratively and simultaneously to obtain the estimates $(\hat{\mu}, \hat{\sigma}, \hat{k})$ of (μ, σ, k) .

Our suggested algorithm is the following:

1. Obtain an initial guess for L .
2. For given L , compute \widehat{k} from $T_L(k)$ using either the bisection method or the Newton Raphson algorithm (we have used the former).
3. Replace the obtained \widehat{k} in $R(\sigma)$ to compute $\widehat{\sigma}$ using the bisection method or the Newton Raphson algorithm (we have used the latter).
4. Substitute \widehat{k} and $\widehat{\sigma}$ into (3) to obtain the corresponding $\widehat{\mu}$.
5. Use these estimates to obtain the next L and to compute the likelihood.
6. Repeat steps (2), (3), (4), and (5) until desired accuracy of the likelihood is achieved.

The iterative procedure was programmed in R and comparison with other optimization methods (i.e. Nelder-Mead and Broyden-Fletcher-Goldfarb-Shanno) offered by software such as R was done using simulations.

2.3 An Approach to Quantile Regression

In reference to Koenker (2005), the p^{th} quantile ($0 < p < 1$) of any real-valued random variable Y is defined as

$$F^{-1}(p) = \inf\{y : F(y) \geq p\},$$

where $F(y)$ is the right continuous distribution function of Y . For a distribution whose density function has the “location-scale” form $\frac{1}{\sigma}f(\frac{y-\mu}{\sigma})$, the p^{th} quantile satisfies

$$F^{-1}(p) = \frac{y - \mu}{\sigma}. \quad (6)$$

By letting the parameters be dependent on a random variable X , i.e. $\mu = a + bx$ and $\sigma = \exp(c + dx)$, we are able to extend the regression model to quantile regression and obtain from (6) quantile curves y_p that explain the data at every p^{th} quantile

$$y_p = a + bx + \exp(c + dx) F^{-1}(p).$$

In the above, we considered the simple “location-scale” example in the context of quantile regression. The generalized gamma has an additional shape parameter, k , that we could make dependent on T and hence extend the model to the 6-parameter case. In our study so far, we analyzed the 4-parameter case, whereby we set μ to be a linear function of T , and similarly to the 3-parameter case, we established an iterative method that estimates the 4 parameters by solving 4 simplified versions of the score equations. We plotted the estimated quantile curves (obtained from \widehat{k} , $\widehat{\sigma}$, \widehat{a} , and \widehat{b}) and compared them to the real ones (obtained from k , σ , a , and b) using simulations.

We aim to extend the estimation of a 4-parameter case to that of a 5 and 6-parameter case. After fitting the 6-parameter case, we perform the likelihood ratio test between the full-parameter model and models with fewer parameters to test the contribution of each parameter in increasing the information about the data. By “elimination”, we conclude whether a 6-parameter model or a smaller one is enough to explain our data. In fact, using a particular distribution, quantile regression, as its name indicates, allows modeling of data at every quantile in contrast to general regression that fits one curve to the whole data set and does not take into account the difference in the behaviour of the data at every quantile.

3 Introducing a New 3-Parameter Distribution

Inspired by the generalized gamma and the need to model skewness, we propose a new unimodal univariate 3-parameter continuous life distribution. While behaviour near zero of the generalized gamma density depends only on α , the tail depends on both shape parameters. Analogously, the basic density function of the new distribution consists of 2 shape parameters; one controls the part next to zero and the other controls the tail. Let θ , α , and $\beta > 0$ be 3 parameters. The density function of the new distribution is

$$f(t) = \frac{\beta}{\theta^\alpha \zeta(\alpha, \beta)} \frac{t^{\alpha-1}}{\left(1 + \frac{t}{\theta}\right)^{\alpha-1}} \exp \left\{ - \left(\frac{t}{\theta}\right)^\beta \right\},$$

where $t > 0$. θ is a scale parameter, α and β are shape parameters, and

$$\zeta(\alpha, \beta) = \int_0^\infty \frac{z^{\frac{\alpha}{\beta}-1}}{\left(1 + z^{1/\beta}\right)^{\alpha-1}} e^{-z} dz.$$

Obviously,

$$\lim_{t \rightarrow 0} f(t) \sim \frac{\left(\frac{t}{\theta}\right)^{\alpha-1}}{\left(1 + \frac{t}{\theta}\right)^{\alpha-1}} \sim \left(\frac{t}{\theta}\right)^{\alpha-1}$$

and

$$\lim_{t \rightarrow \infty} f(t) \sim e^{-\left(\frac{t}{\theta}\right)^\beta} \rightarrow 0.$$

We aim to study the properties of this distribution in terms of maximum likelihood estimation of parameters, and apply similar methods to those used for the generalized gamma, hoping to obtain better estimates.

References

- Abramowitz, M., and Stegun, I. A. (1965). *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. New York: Dover.
- Cox, C., Chu, H., Shneider, M. F., Muñoz, A. (2007). Parametric Survival Analysis and Taxonomy of Hazard Functions for the Generalized Gamma Distribution. *Statistics in Medicine*, **26**, 4352-4374.
- Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.
- Lawless, J. F. (1980). Inference in the Generalized Gamma and Log Gamma Distributions. *Technometrics*, **22**, 409-419.
- Prentice, R. L. (1974). Log-Gamma Model and its Maximum Likelihood Estimation. *Biometrika*, **61**, 539-544.
- Stacy, E. W. (1962). A Generalization of the Gamma Distribution. *Annals of Mathematical Statistics*, **33**, 1187-1192.

An Application of the Multivariate Linear Mixed Model to the Analysis of Shoulder Complexity: EMG Measurements in Breast Cancer Patients

G. Oskrochi¹, E. Lesaffre², M. Molas³ and D. Shamley⁴

¹ School of Technology, Department of Mathematical Sciences, Oxford Brookes University, Oxford, OX33 1HX, UK

² Dept. of Biostatistics, Erasmus University Rotterdam, Dr. Molewaterplein 50, 3015 GE Rotterdam, the Netherlands and Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Katholieke Universiteit Leuven, Kapucijnenvoer 35, Blok D, Bus 7001, 3000 Leuven, Belgium

³ Dept. of Biostatistics, Erasmus University Rotterdam, Dr. Molewaterplein 50, 3015 GE Rotterdam, the Netherlands

⁴ Centre for Postgraduate Medical Research and Education, Bournemouth University, Bournemouth, UK

Keywords: Multivariate Linear Mixed Model; Correlated Random Effects; autoregressive of order one.

1 Introduction

Reduced surgical intervention with irradiation of the remaining breast tissue is now standard procedure for the treatment of breast cancer. Yet despite the use of less extensive surgery there is still morbidity affecting the shoulder. The exact nature of the morbidity and its relationship to pain is not yet known. Radiotherapy has several known effects on parenchyma, vascular tissues and connective tissues. Connective tissue findings suggest that thickening of surrounding connective tissue may restrict movement of related tissues held within the confines of the band. A limited ability to expand together with ischaemia due to changes in the vascular network could have an effect on the efficacy of muscle contraction.

In this study muscle activity was detected by measuring the electrical current generated by muscles in *milli Volts* (EMG) during an upward/downward movement for the affected and the unaffected arm. Four muscles (i.e. pectoralis major, serratus anterior, upper trapezius and Rhomboid) acting on the scapula were investigated in patients treated for breast cancer.

The aims of the study were: (i) to compare shoulder muscle activity for upward/downward arm movement on each side. (ii) to explore any rela-

tionship to the patients report of pain and dysfunction; (iii) to explore any relationship relevant to relevant prognostic factors.

Section 2 introduces data and prognostic factors. Section 3 introduces the multivariate linear mixed models. Section 4 presents the results of the fit of the model to the data. In Section 5 some final conclusions and a discussion of future analyses are given.

2 Data and Prognostic Factors

Two hundred and two patients who had been treated in the last 6 years for unilateral carcinoma of the breast were included in the study. The measurement procedure is detailed in Shamley et al. (2007). Humeral elevation in degrees was measured using the Polhemus FastrakTM motion analysis system and concurrent EMG readings were taken at 10° increments of humeral elevation for each patient. The average of three EMG readings at each elevation point was taken as response (dependent) variable for each patient. All patients filled in a Shoulder Pain and Disability Index (SPADI) questionnaire immediately prior to data collection. The SPADI questionnaire is a known and valid measure of pain and disability for shoulder dysfunction with high levels of sensitivity and reliability (Roach et al., 1991) and is scored on a visual analog scale with 13 items (5 for pain and 8 for disability). Pain scores range from a minimum of 0 to a maximum of 500 and for disability 0 to 800, where 0 representing no symptoms of pain or disability. Possible prognostic factors were: affected side, dominant hand, degree of arm elevation, treatment protocol [Wide Local Excision (WLE) or other], duration after surgery in days, age in year, exercise level, physiotherapy level, receiving chemotherapy, and the Shoulder Pain and Disability Index. The EMG measurements of muscle activity were obtained from four muscles acting at the shoulder (right and left side) of patients. Each muscle activity is recorded in milli Volts (mV) at 10° increments of humeral elevation during upward and downward arm movement. Therefore at each increment point, four EMG measures were obtained, one for each of the four muscles. This creates the multivariate response structure. Further, the fact that the same muscle is measured at different increments creates the 'repeated measures' dimension of the data. Summarized, four repeatedly measured responses for each shoulder and each movement were generated.

3 The Multivariate Linear Mixed Model

Let \mathbf{Y}_{iksm} denote the n_{iksm} -dimensional vector of EMG measurements from the i th patient ($i = 1 \dots N$) on the k th muscle ($k = 1 \dots 4$), during upward ($m = 1$) or downward ($m = 2$) movement of affected arm ($s = 1$) or unaffected arm ($s = 2$). If we ignore all possible associations between

measurements, the analysis can be done assuming a multivariate normal distribution for \mathbf{Y}_{iksm} :

$$\mathbf{Y}_{iksm} \sim \mathcal{N}(\mathbf{X}_{iksm}\beta_{ksm}, \boldsymbol{\Sigma}_k), \tag{1}$$

where \mathbf{Y}_{iksm} is a $n_i \times 1$ vector of EMG readings at different humeral elevation points (n_i is the number of elevation points for individual i), \mathbf{X}_{iksm} is a $n_i \times p$ matrix of covariates (p is the number of covariates), β_{ksm} is a $p \times 1$ vector of coefficients and $\boldsymbol{\Sigma}_k$ is a $n_i \times n_i$ matrix of variance covariance for the k th muscle. Note that we make an assumption that variance covariance matrix depends only on the muscle being analysed, further we assume autoregressive structure of order one for $\boldsymbol{\Sigma}_k$. The AR(1) assumes that n_i observations of \mathbf{Y}_{iksm} are generally correlated, but this correlation is stronger when the humeral elevations are closer and weaker for those observations more apart from each other in term of humeral elevations. Hence $\boldsymbol{\Sigma}_k$ is given by:

$$\boldsymbol{\Sigma}_k = \begin{pmatrix} 1 & \rho_k & \rho_k^2 & \dots & \rho_k^{n_i-1} \\ \sigma_k^2 & \rho_k & \rho_k^2 & \dots & \rho_k^{n_i-1} \\ \rho_k & \rho_k^2 & \rho_k & 1 & \rho_k \\ \vdots & \rho_k & \rho_k & 1 & \rho_k \\ \rho_k^{n_i-1} & \dots & \rho_k^2 & \rho_k & 1 \end{pmatrix} \tag{2}$$

Next, denote \mathbf{Y}_{ik} a vector of all measurements obtained for a muscle:

$$\mathbf{Y}_{ik} = \begin{bmatrix} \mathbf{Y}_{ik11} \\ \mathbf{Y}_{ik12} \\ \mathbf{Y}_{ik21} \\ \mathbf{Y}_{ik22} \end{bmatrix} \sim \mathcal{N} \left[\begin{pmatrix} \mathbf{X}_{ik11}\beta_{k11} \\ \mathbf{X}_{ik12}\beta_{k12} \\ \mathbf{X}_{ik21}\beta_{k21} \\ \mathbf{X}_{ik22}\beta_{k22} \end{pmatrix}, \boldsymbol{\Sigma}_{kk} = \begin{pmatrix} \boldsymbol{\Sigma}_k & 0 & 0 & 0 \\ 0 & \boldsymbol{\Sigma}_k & 0 & 0 \\ 0 & 0 & \boldsymbol{\Sigma}_k & 0 \\ 0 & 0 & 0 & \boldsymbol{\Sigma}_k \end{pmatrix} \right]. \tag{3}$$

Note that in the above model the components \mathbf{Y}_{iksm} are initially assumed to be independent within each muscle and each patient.

It is not realistic to assume that all relevant risk factors or covariates have been measured and included into the statistical model. Unmeasured or omitted risk factors generate a between-case variation often referred to as *frailty* in the biomedical literature. We assume that each muscle may have one such unobserved risk factor, i.e. “muscle specific effect”. Hence all measurements of a muscle are likely to be correlated due to the unobserved muscle’s specific effect. Therefore,

$$\mathbf{Y}_{ik}|v_k \sim \mathcal{N}(\mathbf{X}_{ik}\beta_k + v_{ik}, \boldsymbol{\Sigma}_{kk}) \tag{4}$$

Expression in (4) allows the analysis of the data for muscle k by the means of the univariate linear mixed model with random intercept v_k assumed to follow $\mathcal{N}(0, d_{kk})$ and residual block diagonal variance covariance matrix with blocks $\boldsymbol{\Sigma}_k$ which follows an auto-regressive process of order one.

Since we observe measurements from muscles within patient, therefore we assume that the possible correlation between muscles of the same patient can be modelled via correlations between muscle specific effects.

The joint model is defined by assuming a distribution for the conditional density $\mathbf{Y}_i|\mathbf{v}_i$ and the distribution for multivariate random effects \mathbf{v}_i . These are given as below:

$$\mathbf{Y}_i|(v_{i1}, v_{i2}, v_{i3}, v_{i4}) \sim \mathcal{N} \left[\begin{pmatrix} \mathbf{X}_{i1}\beta_1 + v_{i1} \\ \mathbf{X}_{i2}\beta_2 + v_{i2} \\ \mathbf{X}_{i3}\beta_3 + v_{i3} \\ \mathbf{X}_{i4}\beta_4 + v_{i4} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & 0 & 0 & 0 \\ 0 & \boldsymbol{\Sigma}_{22} & 0 & 0 \\ 0 & 0 & \boldsymbol{\Sigma}_{33} & 0 \\ 0 & 0 & 0 & \boldsymbol{\Sigma}_{44} \end{pmatrix} \right] \quad (5)$$

$$\mathbf{v}_i = \begin{bmatrix} v_{i1} \\ v_{i2} \\ v_{i3} \\ v_{i4} \end{bmatrix} \sim \mathcal{N} \left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{D} = \begin{pmatrix} d_{11} & d_{12} & d_{13} & d_{14} \\ d_{12} & d_{22} & d_{23} & d_{24} \\ d_{13} & d_{23} & d_{33} & d_{34} \\ d_{14} & d_{24} & d_{34} & d_{44} \end{pmatrix} \right]$$

The joint model defined above is a multivariate linear mixed model. Similarly as in the case of a univariate linear mixed model, the marginal likelihood is available analytically serving as an objective function for finding the estimates of β , \mathbf{D} and residual block diagonal variance covariance matrix with blocks $\boldsymbol{\Sigma}_k$ assumed to follow auto-regressive process of order one - AR(1). We used SAS 9.1, PROC MIXED to estimate the model.

4 Empirical Results

Table 1 presents the result of the analysis using model 3, i.e. a multivariate normal distribution for \mathbf{Y}_{iksm} with block diagonal variance covariance matrix with blocks $\boldsymbol{\Sigma}_k$. The likelihood ratio test suggests that the effect of arm movement (MOVE) is best presented by a dummy variable indicating different intercepts for the upward and downward movement. Statistical tests suggest that the effect of affected shoulder and unaffected shoulder cannot be modelled by considering only different intercepts for affected/unaffected shoulder. This implies that the effect of prognostic factors is different on the affected shoulder as to the unaffected shoulder; therefore interaction model was employed to assess the interaction effects of shoulders on all prognostic factors.

The values in bold correspond to a significant (at the 5% level) result for each muscle. In the column ‘overall significance (O. Sig)’ it is shown whether the prognostic factor is collectively (for 4 degrees of freedom) significant. The final column (In. Eff) shows whether interaction effect is collectively significant or not. This model suggests that humeral elevation of the arm

TABLE 1. Result of the analysis using a multivariate normal distribution with AR(1).

Variables	Ln(PM)	Ln(UT)	Ln(SA)	Ln(RH)	O. Sig	In. Eff
Const	2.91	3.35	2.82	2.16	< 0.001	N/A
Degree	0.50	0.59	0.84	0.71	< 0.001	No
Move	0.11	0.29	0.12	0.15	< 0.001	No
Affect	-0.12	-0.53	-0.21	-0.22	0.800	N/A
Hand	0.41	0.11	0.16	0.16	0.002	No
Domin	0.17	0.07	0.06	0.07	0.552	Yes
Age $\times 10^3$	-1.64	2.38	-6.14	3.07	0.406	No
Durat $\times 10^3$	-0.10	0.04	0.14	0.15	0.008	No
Sppain $\times 10^3$	-1.67	-2.15	-1.54	-1.65	< 0.001	Yes
Spdis $\times 10^3$	0.86	1.07	1.46	0.04	0.048	No
Treatm	-0.22	-0.27	-0.29	-0.24	< 0.001	No
Chemo	0.04	-0.14	-0.58	-0.21	< 0.001	No
Exer AS	0.00	-0.03	-0.01	0.00	0.828	No
Exer. C	-0.01	0.04	0.05	0.05	0.003	Yes
Physio1	0.37	0.02	0.02	-0.38	0.328	No
Physio2	-0.03	0.13	0.13	0.13	0.348	No

(degree), the upwards move, left hand, duration since surgery and currently doing exercise (Exer. C) will increase the electrical activity generally for all four muscles irrespective of which side is affected; while patients treated with wide local excision, chemotherapy and SPADI pain have decreased electrical activity on both sides. Interaction analysis suggests that the effect of affected or unaffected shoulder is significantly different on dominant hand, SPADI pain, and current exercise. The deviance ($-2 \times \log$ likelihood) for this model is 29138.4. The disadvantage of model (3) is that it ignores any possible associations between measurements.

The following table is the result of fitting a multivariate linear mixed model (5) with correlated random intercept v_k and residual block diagonal variance covariance matrix with blocks Σ_k assumed to follow auto-regressive process of order one.

In Table 2 we show the results of fitting model (5) to the data. Similar to the results of the previous model, this model is also now suggested that humeral elevation of the arm (degree) and the upward movement will increase the electrical activity of all four muscles irrespective of which side is affected. Duration since surgery is significant but it only affects the affected arm. Again treatment with wide local excision, and receiving chemotherapy decreases the electrical activity of all muscles. Interestingly, the left/right shoulder (hand), SPADI pain, SPADI disability and exercise now are no

TABLE 2. Result of the analysis using Multivariate Linear Mixed Model with AR(1).

Variables	Ln(PM)	Ln(UT)	Ln(SA)	Ln(RH)	O. Sig	In. Eff
Const	2.99	3.38	2.83	2.18	< 0.001	N/A
Degree	0.50	0.57	0.83	0.69	< 0.001	No
Move	0.13	0.30	0.13	0.19	< 0.001	No
Affect	-0.01	-0.39	0.37	-0.04	0.680	No
Hand	0.40	0.10	-0.17	0.06	0.137	N/A
Domin	0.17	0.06	0.06	0.26	0.885	No
Age $\times 10^3$	-2.07	2.60	-5.94	3.16	0.381	No
Durat $\times 10^3$	-0.11	0.04	0.13	0.14	0.007	Yes
Sppain $\times 10^3$	-1.70	-2.14	-1.52	-1.72	0.230	Yes
Spdis $\times 10^3$	0.81	1.02	1.33	0.07	0.378	No
Treatm	-0.22	-0.26	-0.27	-0.22	0.048	No
Chemo	0.03	-0.12	-0.56	-0.21	0.002	Yes
Exer. AS	-0.00	-0.04	-0.02	-0.00	0.813	Yes
Exer. C	-0.01	0.04	0.05	0.05	0.057	Yes
Physio1	0.37	0.04	0.03	-0.37	0.239	Yes
Physio2	-0.04	0.11	0.13	0.12	0.586	Yes

longer significant irrespective of the affected shoulder. But, the interaction analysis suggests that the effect of affected or unaffected shoulder is significantly different on duration, SPADI pain, chemotherapy, exercise 6 months after surgery, current exercise, physiotherapy after surgery and current physiotherapy. The deviance ($-2 \times \log$ likelihood) for this model is 26776.0 and gain in $-2 \times \log$ -likelihood is 2362.4 for 10 degrees of freedom. Hence, using a simple conventional AR(1) model for parameter estimates for each muscle and ignoring the existing associations between the measurements leads to unrealistic inferences for important diagnostic factors. Applying model (4) while ignoring the associations between muscle specific effects leads to clinically more sensible parameter estimates and substantial improvement in term of the likelihood ratio test; a deviance difference of 2007.9 for 4 degrees of freedom.

Comparing the likelihoods of model (1) and (4) confirms the presence of strong muscle specific effects.

Model (5) assesses the presence of significant association between muscle specific random effects. A proper comparison between the likelihood of model (4) and (5) returns a deviance difference of 331.7 for 6 degrees of freedom. This suggests the use of a joint multivariate linear mixed model with residual block diagonal variance covariance matrix which assumed to follow auto-regressive process of order one, is more appropriate.

5 Discussion

Irrespective of side affected, this study has shown:

- Increase of activity in four key muscles acting on the shoulder complex during elevation of the arm (the higher is the elevation of the arm the higher is the EMG activity in all muscles).
- Significant loss of all muscle activity on the downward movement as compared to upward movement of the arm indicates loss of eccentric muscle control of the shoulder girdle against gravity. The significance of this has to be tested in a case-control type of study which includes healthy individuals.
- Collectively, muscle activity is statistically not different on the left shoulder of the patients as compare to the right shoulder.
- Muscle activity failed to show any relation to hand dominance.
- Higher pain score is significantly associated with a drop in UT activity, but not with any other muscle activity.
- Reduced activity in all four muscles is associated with the WLE treatment group.
- While chemotherapy is generally associated with a drop in muscle activity, it affects SA more than the others.
- Current exercise, exercise after surgery and physiotherapy has no significant effect on muscle activity.

For the affected shoulder this study has shown:

- Duration since surgery is associated with drop in muscle activity in SA and RH.
- SPADI pain is positively associated with increase in muscle activity.
- Receiving chemotherapy is significantly associated with drop in PM activity of the affected shoulder.
- Current exercise is associated with increase in muscle activity for PM, UT and SA.
- Physiotherapy is also associated with increasing muscle activity in UT and SA.

These conclusions are interesting from a clinical perspective since, during elevation of the arm UT and SA work as a force couple to protract, elevate and laterally rotate the scapula thereby ensuring clearance of the subacromial arch. The loss of UT and SA activity would alter the force couple produced by these muscles, placing the scapula in a protracted and medially rotated position, thereby decreasing upward movement; interestingly current exercise significantly improves UT and SA activity.

References

- Roach KE, Budiman-Mak E, Songsriridej N, Lertatanakul Y (1991). Development of a shoulder pain and disability index. *Arthritis Care and Research*; 4: 143-49.
- SAS Institute Inc. 100 SAS Campus Drive Cary, NC 27513-2414, USA
- Shamley, D., Srinanaganathan R. , Weatherall R., and Oskrochi R. et. al. (2008). Changes in shoulder muscle size and activity following treatment for breast cancer. *Journal of Breast Cancer Research and treatments*.
- Williams, JW, Holleman, DR and Simel, DL (1995). Measuring shoulder function with the shoulder pain and disability index. *J Rheumatol Apr*; 22(4):727-32.

A test procedure for right censored data under the additive model

Hyo-Il Park¹ and Seung-Man Hong²

¹ Department of Statistics, Chongju University, Chongju, Choong-book 360-764, Korea email:hipark@cju.ac.kr

² Department of Informational Statistics, Korea University, Jochiwon, Choongnam 339-700, Korea email:smhong@korea.ac.kr

Abstract: In this research, we propose a nonparametric test procedure for the right censored and grouped data under the additive hazards model. For deriving the test statistics, we use the likelihood principle. Then we illustrate proposed test with an example. Finally we discuss some interesting features concerning the proposed test.

Keywords: log-rank statistic; score function.

1 Introduction

The proportional hazards model (PHM) has been one of the most frequently applied ones for the analysis of the life-time data. Since Cox(1972) has proposed the PHM, the PHM has been developed and modified successfully in many various situations. However when the proportionality among hazard functions may be suspicious, one may as well consider an alternative model rather than clinging to the PHM. Then the additive hazards model (AHM) may be a candidate for any possible alternatives. For $t \in [0, \infty)$ let $\lambda_0(t)$ be the baseline hazard function and $z = (z_1, \dots, z_p)'$, the $p \times 1$ regression vector, where the prime represents the transpose of a vector or matrix. Then the hazard function $\lambda(t, z)$ for the AHM can be represented with the $p \times 1$ regression coefficient vector $\beta = (\beta_1, \dots, \beta_p)'$ as follows:

$$\lambda(t, z) = \lambda_0(t) + \beta'z. \quad (1.1)$$

Then the corresponding cumulative hazard function, $\Lambda(t, z)$ and survival function, $S(t, z)$ under the AHM (1.1) can be written as follows with the facts that $\int_0^t \beta'z dx$ and $S(t) = \exp[-\Lambda(t)]$:

$$\Lambda(t, z) = \int_0^t (\lambda_0(x) + \beta'z) dx = \Lambda_0(t) + t\beta'z$$

and

$$S(t, z) = \exp[-\Lambda_0(t)] \exp[-t\beta'z]. \quad (1.2)$$

As an alternative model to the PHM, the AHM has not been widely used. The main reason for this may come from the fact that the conditional likelihood proposed by Cox (1972) can not be applied to the AHM because of the structure of the hazard function. The AHM (1.1) was initiated by Aalen(1980, 1989), who considered an inference procedure for λ_0 and β applying the least squares method. McKeague(1988) and Huffer and McKeague(1991) considered the weighted least squares estimates for some optimality consideration. Also Lin and Ying(1994) proposed an estimate procedure for β using the counting process which has been used for the PHM as an ad hoc approach. McKeague and Sasieni(1994) developed partly parametric AHM. Also Scheike(2002) worked the AHM in this direction. For the multivariate data, Yin and Cai(2004) considered inferences based on the marginal AHM approach.

Sometimes one cannot help observing the objects whether they fail or not periodically or with time-schedule for some reasons. For example, after being exposed to the HIV virus, the observation must be carried out periodically since it usually takes several months for blood test results from HIV negative to HIV positive. In this case data set contains lots of tied value observations even though the underlying life-time distribution is continuous. This type of data set is called the grouped data and should be analyzed by a data-specific method. Heitjan(1989) reviewed extensively the methodology and suggested several research directions. For the right censored data, Prentice and Gloeckler(1978) considered the inferences about β under the PHM. Park(1993) proposed a class of nonparametric tests for the linear model whereas Neuhaus(1993) modified the so-called log-rank tests for the grouped data. In this study, we consider to propose a nonparametric test for β under the AHM (1.1) using the score function based on the likelihood principle for the grouped and right censored data. The scores will be derived using the discrete model approach (cf. Kalbfleisch and Prentice, 1980). Then we illustrate our test with an example. Finally we discuss some interesting features about our test procedure.

2 Nonparametric test

Suppose that we observe life time T_i for the i th individual with some $p \times 1$ covariate vector, $z_i = (z_{i1}, \dots, z_{ip})'$, $i = 1, \dots, n$. We assume that each subject is prone to be censored. In this way, the data set can be represented as $\{(T_i, \delta_i, z_i), i = 1, \dots, n\}$, where δ_i stands for the censoring status with values 0 or 1 if censored or not. We assume that the hazard function for each individual follows AHM (1.1) but do not assume any specific form for the baseline hazard function, λ_0 . Also we assume that survival and censoring random variables are independent and distributions of censoring random variables do not contain any information about β . Since we are concerned with the grouped data, we assume that the positive half real line, $[0, \infty)$ is

partitioned into k sub-intervals such as $[0, \infty) = \bigcup_{l=1}^k [a_{l-1}, a_l)$, with $a_0 = 0$ and $a_k = \infty$. Then one can only have the information that T_i is contained in one of k sub-intervals for all i . We denote D_l and C_l as the indicate sets for the uncensored and censored observations in the l th sub-interval $[a_{l-1}, a_l)$, respectively. Also we denote R_l as the risk set of the l th sub-interval. Finally we denote d_l and r_l as the sizes of D_l and R_l , respectively, $l = 1, \dots, k$. In this grouped continuous data, we assume that all the censorings occur at the end of a sub-interval and all the deaths precede any censoring in the same sub-interval. Finally we assume that all the observations in the last sub-interval $[a_{k-1}, \infty)$ are censored at a_{k-1} for some technical reason. Then from the discrete model in Kalbfleisch and Prentice(1980) with all the assumptions and notation introduced up to now, we have with (1.2) that for each $l, l = 1, \dots, k - 1$

$$\Pr \{T_i \in [a_{l-1}, a_l), \delta_i = 1, z_i\} \propto \exp [-\Lambda_0(a_{l-1})] \exp [-a_{l-1}\beta' z_i] - \exp [-\Lambda_0(a_l)] \exp [-a_l\beta' z_i]$$

and

$$\Pr \{T_i \in [a_{l-1}, a_l), \delta_i = 0, z_i\} \propto \exp [-\Lambda_0(a_l)] \exp [-a_l\beta' z_i].$$

For $l = k$, we have that

$$\Pr \{T_i \in [a_{k-1}, \infty), \delta_i = 0, z_i\} \propto \exp [-\Lambda_0(a_{k-1})] \exp [-a_{k-1}\beta' z_i].$$

Thus under AHM (1.1), the likelihood function for the discrete model becomes as

$$L(\beta) \propto \prod_{l=1}^{k-1} \left\{ \prod_{i \in D_l} (\exp [\Lambda_0(a_l) - \Lambda_0(a_{l-1})] \exp [(a_l - a_{l-1})\beta' z_i] - 1) \prod_{i \in D_l \cup C_l} \exp [-\Lambda_0(a_l)] \exp [-a_l\beta' z_i] \right\} \prod_{i \in C_k} \exp [-\Lambda_0(a_{k-1})] \exp [-a_{k-1}\beta' z_i].$$

Then by taking logarithm on $L(\beta)$, differentiating partially the log-likelihood function with respect to β_j , substituting 0 for β_j , and re-arranging the result with some algebraic manipulation for each $j, j = 1, \dots, p$, we have that

$$W_{jn} = \sum_{l=1}^{k-1} (a_l - a_{l-1}) \frac{r_l}{d_l} \left\{ \sum_{i \in D_l} z_{ij} - \frac{d_l}{r_l} \sum_{i \in R_l} z_{ij} \right\}.$$

Then we note that W_{jn} is a score statistic and can be used for testing $H_0^j : \beta_j = 0$. Then for the construction of a test statistic for testing $H_0 : \beta = 0$, we need the means and variances of $W_{jn}, j = 1, \dots, p$ and covariances between W_{jn} and $W_{j'n}$ for $j \neq j'$ under $H_0 : \beta = 0$. For this, first of all,

we note that W_{jn} is a martingale with discrete compensators (Flemming and Harrington, 1991). From this fact, one may easily conclude that under $H_0 : \beta = 0$ the mean of W_{jn} is 0 and an unbiased estimate $\hat{\sigma}_{jn}^2$ of the variance of W_{jn} is of the form

$$\hat{\sigma}_{jn}^2 = \sum_{l=1}^{k-1} (a_l - a_{l-1})^2 \frac{r_l(r_l - d_l)}{(r_l - 1)d_l^2} \sum_{i \in R_l} (z_{ij} - \bar{z}_{lj})^2,$$

where $\bar{z}_{lj} = (1/r_l) \sum_{i \in R_l} z_{ij}$. Also an unbiased null covariance estimate $\hat{\sigma}_{jj'n}$ of the covariance between W_{jn} and $W_{j'n}$ for $j \neq j'$ can be obtained by the same arguments used for the null variance estimate by noticing that the covariance between observations with z_{ij} and $z_{i'j'}$ is 0 whenever $i \neq i'$. Thus an unbiased estimate, $\hat{\sigma}_{jj'n}$ becomes of the form

$$\hat{\sigma}_{jj'n} = \sum_{l=1}^{k-1} (a_l - a_{l-1})^2 \frac{r_l(r_l - d_l)}{(r_l - 1)d_l^2} \sum_{i \in R_l} (z_{ij} - \bar{z}_{lj})(z_{i'j'} - \bar{z}_{lj'}),$$

where $\bar{z}_{lj'}$ can be defined similarly with \bar{z}_{lj} . Let $\hat{V}_n = (\hat{\sigma}_{jj'n})_{j,j'=1,\dots,p}$ with $\hat{\sigma}_{jjn} = \hat{\sigma}_{jn}^2$. Then with the assumption that \hat{V}_n is nonsingular, one may propose the following quadratic form for a test statistic for testing $H_0 : \beta = 0$

$$Q_n = \begin{pmatrix} W_{1n} \\ \vdots \\ W_{pn} \end{pmatrix}' \hat{V}_n^{-1} \begin{pmatrix} W_{1n} \\ \vdots \\ W_{pn} \end{pmatrix},$$

where \hat{V}_n^{-1} is the inverse of \hat{V}_n . Then one may reject $H_0 : \beta = 0$ in favor of $H_1 : \beta \neq 0$ for large values of Q_n . In order to have critical value for any given significance level, we need the null distribution of Q_n . Since the null distribution of Q_n depends on the unknown censoring distribution, we consider to obtain the limiting distribution of Q_n . For this first of all, we need the following lemma.

Lemma. For each $j, j = 1, \dots, p$, under all the assumptions used up to now and with the following condition that

$$\max \frac{1}{\sqrt{n}} \{z_{1j}, \dots, z_{nj}\} \rightarrow 0, \tag{2.1}$$

we have that under $H_0 : \beta = 0$

$$W_{jn} / \sqrt{\hat{\sigma}_{jn}^2}$$

converges in distribution to a standard normal distribution as $n \rightarrow \infty$.

Then with all the notation introduced up to now, we state the following main result.

Theorem. With the assumption that \hat{V}_n is nonsingular and the condition (2.1) for each $j, j = 1, \dots, p$, under $H_0 : \beta = 0$, Q_n converges to a chi-square distribution with p degrees of freedom.

Proof. From Lemma, the Cramér-Wold device (cf. Billingsley, 1986) and the Slutsky's theorem with the assumptions that \hat{V}_n is a nonsingular consistent estimate and (2.1), the result follows easily.

3 An example and some concluding remarks

In order to illustrate our test procedure, we consider the data reported by Embury et al.(1977) for the length of remission (in weeks) with acute myelogenous leukemia patients. The data have been summarized as follows:

Control group: 5 5 8 8 12 16+ 23 27 30 33 43 45
 Maintenance group: 9 13 13+ 18 23 28+ 31 34 45+ 48 161+.

+ indicates censored observation. Since the length of remission for each patient was measured by week, the data set contains several tied observations and sub-interval may be designated by each week. Thus the lengths of sub-intervals are all the same with unity. From Figure 1, assuming model (1.1) may be appropriate since the graphs of two cumulative hazards seem to be parallel each other. The object of this experiment was to see if the maintenance chemotherapy prolongs the length of remission. Therefore we consider to use W_n instead of Q_n since we should consider one-sided alternative $H_1 : \beta > 0$. Then by allocating 0 or 1 to covariate z_i for the i th individual according as control or maintenance chemotherapy group, we have that

$$W_{19} = 27.5 \text{ and } \hat{\sigma}_{19}^2 = 253.5183.$$

Since $W_{19}/\sqrt{\hat{\sigma}_{19}^2} = 1.73$, we obtain 0.042 as its p -value from the table of a standard normal distribution. In passing, we note that the procedure proposed by Prentice and Gloeckler(1978) gives 0.065 as its p -value.

For now we discuss some interesting feature about our proposed test statistic. For this, we consider the case that $p = 1$. Also we consider W_n itself rather than the quadratic form Q_n . Then under the two-sample problem setting, W_n can be re-written as

$$W_n = \sum_{l=1}^{k-1} (a_l - a_{l-1}) r_{1l} r_{2l} \left\{ \frac{d_{1l}}{r_{1l}} - \frac{d_{2l}}{r_{2l}} \right\}, \tag{3.1}$$

where d_{jl} and r_{jl} are the sizes of the deaths and risk set in the l th sub-interval for the j th sample, $j = 1, 2$. We note that W_n in (3.1) is the Gehan statistic(cf. Gill, 1980) apart from the length of sub-interval. Therefore one may consider that W_n is an extension of the Gehan statistic for the

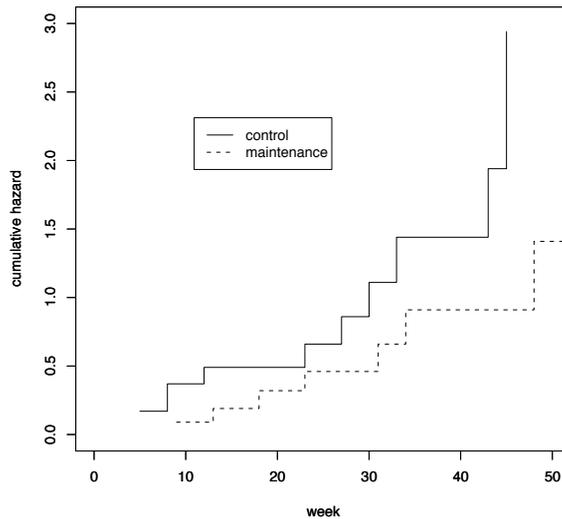


FIGURE 1. Cumulative Hazard Plots for Leukemia Data.

grouped case under AHM (1.1). Since the Gehan test is a modification of the Wilcoxon test for the censored observations, the test based on W_n may be optimal in the sense of power for the location translation alternatives. We could confirm this fact from a simulation study, which is not included in this presentation.

When deriving test statistic, we assumed that all the observations in the last sub-interval $[a_{k-1}, \infty)$ are censored at a_{k-1} , the beginning point of the last sub-interval. The reason for this is as follows. First of all, we note that the length of the last sub-interval is infinity. If there is any uncensored observation in the last sub-interval, then the length of the last sub-interval should be included in W_{jn} , which is an absurd expression. Also if we maintain the assumption that the censoring occurs at the end of each sub-interval even to the last sub-interval, then the derivation of W_{jn} becomes impossible for the censored observations in the last sub-interval. However in the real experimental design, since always a researcher observes objects during a finite time period, such assumption becomes insignificant and can be avoided being required such restriction by adding more sub-intervals.

For the null distribution, we derived asymptotic normality using the large sample approximation. Also one may consider a re-sampling approach such as the permutation principle (cf. Good, 2000) to obtain a null distribution.

Park(1993) and Neuhaus(1993) considered to apply the permutation principle for the right censored and grouped data under some different models. When one applies the permutation principle for the censored data, one must include the condition of equality of unknown censoring distributions in the null hypothesis. The resulting permutation test is known to be exact but conditional. Also as another re-sampling method, one may use the bootstrap method. For the censored data, you may refer to Efron(1981) and Reid(1981). Unlike the permutation principle, the bootstrap method does not require equality among censoring distributions. Because of the computational amount of work, application of re-sampling methods takes in general the Monte-Carlo approach.

Acknowledgments: This research was supported by the Korea Science and Engineering Foundation grant funded by the Korea government(MOST). (No. 2009-0052725)

References

- Aalen, O.O. (1980). A model for non-parametric regression analysis of counting processes. *Springer Lecture Notes Statistics 2*, 1-25. *Mathematical Statistics and Probability Theory*, W. Klonecki, A. Kozek and J. Rosinski, editors.
- Aalen, O.O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, **8**, 907-925.
- Billingsley, P. (1986). *Probability and Measure*, Second Edition. New York: Wiley & Sons.
- Cox, D.R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 189-220.
- Efron, B.R. (1981). Censored data and the bootstrap. *Journal of American Statistical Association*, **76**, 312-319.
- Embury, S.H., Elias, L., Heller, P.H., Hood, C.E., Greenberg, P.L. and Schrier, S.L. (1977). Remission maintenance therapy in acute myelogenous leukemia. *Western Journal of Medicine*, **126**, 267-272.
- Flemming, T.R. and Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. New York: Wiley & Sons.
- Gill, R.D. (1980). *Censoring and Stochastic Integrals*. Amsterdam:Mathematical Centre Tracts, Mathematisch Centrum.
- Good, P. (2000). *Permutation Tests-A Practical Guide to Resampling Methods for Testing Hypothesis*, Second Edition. New York: Springer.

- Heitjan, D.F. (1989). Grouped continuous data. *Statistical Science*, **4**, 164-183.
- Huffer, F.W. and McKeague, I.W. (1991). Weighted test squares estimation for Aalen's additive risk model. *Journal of American Statistical Association*, **86**, 114-129.
- Jones, M.P. and Crowley, J. (1990). Asymptotic properties of a generalized class of nonparametric tests for survival analysis. *Annals of Statistics*, **18**, 1203-1220.
- Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley & Sons.
- Lin, D.Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, **81**, 61-71.
- McKeague, I.W. (1988). A counting process approach to the regression analysis of grouped survival data. *Stochastic Process with Applications*, **28**, 221-239.
- McKeague, I.W. and Sasieni, P.D. (1994). A partly parametric additive risk model. *Biometrika*, **81**, 501-514.
- Neuhaus, G. (1993). Conditional rank tests for the two-sample problem under random censorship. *Annals of Statistics*, **21**, 1760-1779.
- Park, H.-I. (1993). Nonparametric rank-order tests for the right censored and grouped data in linear model. *Communication in Statistics-Theory and Methods*, **22**, 3143-3158.
- Prentice, R.L. and Gloeckler, L.A. (1978). Regression analysis of grouped data with applications to breast cancer data. *Biometrics*, **34**, 57-67.
- Reid, N. (1981). Estimating median survival time. *Biometrika*, **68**, 601-608.
- Scheike, T.H. (2002). The additive nonparametric and semiparametric Aalen model as the rate function for a counting process. *Lifetime Data Analysis*, **8**, 247-262.
- Yin, G. and Cai, J. (2004). Additive hazards model with multivariate failure time data. *Biometrika*, **91**, 801-818.

Computation of Agriculture

Chancal Pramanik¹

¹ [M.Sc. (Ag.) Statistics] Room No.-4; A-Block, P.G. Hostel (Haritha Nilayam), College of Agriculture, Acharya N. G. Ranga Agricultural University, Rajendranagar, Hyderabad-500030, State- Andhra Pradesh, Country-India, E-mail ID- cpramanik@gmail.com

Abstract: India is an Agriculture based country. About 95 percent of hers Agriculture is unorganized. Here is the suggestion how Agriculture can be organized in India, thus improving hers economic status. Computation of Agriculture can be used as a tool to organize Agriculture.

Keywords: Agriculture in India; Computation of Agriculture; Methodology; A dream project; Possible benefits .

1 Introduction

India the Subcontinent is known for hers Unity in Diversity from the ancient past. Diversifications exist in every aspects of this vast country makes her really beautiful that welcomes everyone heartily throughout the ages. Many people from many countries have found their sweet home on this motherland. Thus contributed and contributing diversification in religions, languages, nature of livings, food habits etc. India is basically an Agriculture based country where most of the people are engaged in Farming and Livestock raising. Thus economic condition of India i.e., of hers people vastly depends on Agriculture without any doubt. It is to keep in mind that about 70 percent of Indian rural people depend on agriculture which comprises the major population.

1.1 Agriculture in India today

India ranks second worldwide in farm output. Agriculture and allied sectors accounted for 17.8 percent of the Growth Domestic Product (GDP) in 2007 (according to World Bank report), employed 60 percent of the total workforce. It is the largest economic sector and plays a significant role in the overall socio-economic development of India. Agriculture is also the single largest source of employment in India, even though its contribution to the national economy has been shrinking over the years. Thus Indian Economy vastly depends on Agriculture and allied sectors. So improving agriculture

status in India can improve the Indian Economy. Again, 5 percent of Indian agriculture is organized. Hence most of the agriculture based people in India are employed in an unorganized sector which comprises the major population. With the growth of population food security is the major problem not only in India but throughout the world. Now organizing Agriculture with the help of Statistical Modeling system can serve the world better by optimizing the natural resources.

1.2 Computation of Agriculture and its need

Agriculture is a vast domain comprising of multiple factor interaction. There are many interlinks between various factors of the environment on which agriculture depends on. Besides the nature, agriculture is also depending on social, political and economical status of a region. Hence agricultural output is influenced by many parameters. So it is very much difficult for an individual farmer or an organization to take correct decision about the judicial inputs, management required for his farmland. Thus, agricultural systems need an experienced system to suggest it judicially for its welfare. Here is the need for Statistical Modeling of the agril-based technologies that can act both as a decision support system and as an expert support system which integrates the Computation of Agriculture.

2 Methodologies

Computation of Agriculture need to be followed by certain well-disciplined methodologies. Here are some general methodoliges that can be followed while modeling certain agricultural systems. They are as follows in brief:

I. Define the problem

1. Divide the problem into a series of specific questions.
2. Discuss the problem in a qualitative fashion and present heuristic motivation before plunging into equations.
3. Consciously decide among alternative approaches.
4. State what results we anticipate. This is equivalent to a hypothesis.
5. Review the relevant literature.

II. Plan its treatment

1. Review the relevant physical or chemical principles. Express them verbally and then mathematically.
2. Identify the assumptions that may be desirable and /or necessary.
3. Formulate the problem mathematically.
 - i. Use sketches freely*
 - ii. Define all symbols and give the associated units.*
 - iii. Show all coordinates with positive directions identified.*
 - iv. Identify all assumptions as they are embedded in the model.*
 - v. Cite all references consulted.*

III. Execute the plan

1. Use more than one approach possible.
2. When carrying out the mathematical steps, number all equations.
3. Use frequent intermediate checks.
4. Provide an account of the motivation for each step, and indicate the logical connection between steps.
5. If an impasse is reached, simplify the problem by altering the assumptions and / or the problem definition.

IV. Check thoroughly

1. Check dimensions.
2. Check limiting cases.
3. Check symmetry.
4. Check reasonableness.
5. Check range of variables.
6. Re-examine all assumptions for probable effect and importance.
7. Use alternative derivation.
8. Use analogies.

V. Learn and generalize from the analysis

1. Summarize the important findings, including limitations.
2. Interpret the equations in terms of the original problem.
3. Have important factors been overlooked?
4. What is importance by its absence?
5. Use dimensionless numbers in the graphical presentation.
6. Verbally state the meaning of the final mathematical equations.
7. Can the proposed scheme be improved or altered so that it will not work?
8. Check numerical cases for clarity and feasibility. This is a check of the magnitude of the effect and the required parameter values.
9. How much parameter variation can be tolerated?
10. How many significant figures should be used?
11. Is the graphical presentation fully clarified, including clearly labeled axes, and is an interpretation of the information presented?
12. What uncertainties still exist?
13. How does the model relate to the original problem?
14. Can and should the problem now be simplified or generalized?

After constructing the model it should be programmed in software for its efficient application in the desired field. The following points can be kept in mind while software making:

1. Make needs assessments : Determine from discussions with the proposed clients that there is a real and economically viable need for program development and that there is a chance of meeting that need. Many programs are produced only because there is a body of knowledge that can be expressed.

2. Involve users : Both users of programs and users of the information generated should be included on developmental teams from the beginning, not simply to follow progress and to evaluate output.

3. Clearly define important problems : Program logic needs to be expressed from the perspective of the people who have the problem. Problem definition must be from an information users viewpoint and not be limited by the domain specialist.

4. Not worry about software and hardware : Developers were advised to leave software writing to specialists writing and consider program needs, economics, and availability when selecting hardware.

5. Address the need of potential users : Users may be farmers, extension service specialists, scientists, agribusiness people or other clients. They seek information and interpretations. To be understood, outputs should be in the language of information users, not that of domain experts.

6. Allow for maintenance : The 80/20 rule characterizes the development process. It takes 20 percent of the effort to develop a system or program, and other 80 percent for validation, delivery, and maintenance.

7. Know the benefactors : Keep mind the people who are funding the work.

3 A Project on Computation of Agriculture

Thus on developing the efficient statistical model it will act as a priest to give marriage with the ever growing Information Technology (IT) sector and Agriculture Sector by making software technologies based on the models. This will be the principal theme of my dream project. Combining both the above said sectors can employ more efficient technology based person in Agriculture Sector following up extensive real field research that is the most wanted thing now in this Sector. A group can be made comprising scientists from various disciplines of agriculture, software engineer with chairmanship of Agricultural Statistician including region basis representative from farmers community. The group will help to build an efficient database on the basis of which models can be constructed with more efficiency. The system should be supported with internet facilities, Global Positioning System that can give the models correct information about weather conditions, soil conditions possible pests attack and other related informations related to agricultural development. The system can also support the farmers in decision making in the field level in vernacular language. Thus the integrated system will be able to help the farmer to optimize his available resources. These combined efforts can grow interest in IT sector to build new Agriculture based Software Technology with emphases on region specific. This will introduce more funds in agriculture sector from both private and government organizations. It will create a competition in Agriculture based Software market to do and serve better which will optimize the agricultural resources and steady growth in agriculture sector can be firmly predicted. It will result organization of Agriculture. Hence a new era in agriculture sector can be introduced which can be shortly defined as **Computation of Agriculture** or **e-agriculture**.

3.1 Merits

However the entire system of Computation of Agriculture will likely to bring Second Green Revolution in India, according to my opinion, can have the following merits: 1. Increasing both the production and productivity of the farm products by decreasing the risk of food security problem. 2. Optimizing the natural resources judiciously. 3. Preventing pre harvest and post harvest crop losses. 4. Improving the contribution of agriculture in GDP, thus increasing the condition of the farmers. 5. Developing good forecasting system. 6. Developing good databases on Agricultural Systems. 7. Involving Information Technology (IT) in Agriculture vastly. 9. Introducing more funds and interests in Agricultural Systems, inducing more field research. 10. Boosting the IT sector by providing a big project to run.

3.2 Demerits

1. Initial cost to install the entire system is high. 2. Skilled persons will be required to maintain the entire system. 3. False prediction may results to huge loss.

3.3 The Final Comment

From the above discussions I think it is clear about my dream project Computation of Agriculture to my dear readers. I do strongly believe that if this project can be run efficiently in safe hands, it will be a great success. The entire agriculture system in India will be changed. There is a good chance to bring Second Green Revolution in India by the soft hands of Computation of Agriculture. Now its an appeal to the world to give suggestions and contributions to the project, so that it can be developed efficiently. Of course, the entire system is an integrated process of various conditions and it is to be adopted with passage of time. But if it can be properly make into run it will be a great boon for us, not only for the Indians but also to the world . Hence I am looking forward to introduce a new era in the Indian Agriculture i.e., the **Computation of Agriculture**.

Acknowledgments: Special thanks to my mother Mrs. Debjani Pramanik, my Rangamama Prof. Suhrit K. Dey, my Notunmama Mr. Parimal Kanti Das, my Chotomama Mr. Surajit Das, my Boromama Mr. C. R. Das, Mr. Chandan Das, my friends Mr. Ankon Sen and Mr. Sumit Mukherjee, Prof. James G. Booth, Prof. M. S. Swaminathan, Prof. Sisir K. Mukhopadhyay, Prof. B. S. Kulkarni, Dr. A. Pratap Kumar Reddy, Prof. V. Krishna Rao, Mr. V. V. Narendranath, Mr. Prithwiraj Deb and to all of my well wishers and friends. I do offer every success of mine to the holy Lotus Feet of my Lord Shree Ramji...

References

- Barett John R., Mark A. Nearing. (1998). *Agricultural Systems Modeling and Simulations*. Marcel Dekker, Inc.
- Cooke J. Robert. (1998). *Agricultural Systems Modeling and Simulations*. Marcel Dekker, Inc.
- Das M. N. (December,1999). Technical Address: Use of Statistics and Computer for Agricultural Research and Development. *Journal of the Indian Society of Agricultural Statistics*, **52(3)**, 257-261.
- Johannesson G., Jeffrey Stewart, Liz Brady Sabeff, Dona Heimiller, Anellia Milbrandt. (January 27, 2006). Spatial Statistical procedures to validate Input data in Energy Models. In: <http://www.nrel.gov/docs/fy06osti/39497.pdf>, 1-38.
- Kulkarni B. S., M. Narayana Reddy, K.V. Kumar. (November 2008). Software for Agricultural Research and Education. In: *Souvenir 62nd Annual Conference of Indian Society of Agricultural Statistics (ISAS)*. 20-27, Tirupati, Andhra Pradesh.
- Narendranath V.V., Chanchal Pramanik. (November 2008). Biometrics and its Application in Agriculture. In: *Souvenir 62nd Annual Conference of Indian Society of Agricultural Statistics (ISAS)*. 43-46, Tirupati, Andhra Pradesh.
- Panesar B. S. (1998). *Agricultural Systems Modeling and Simulations*. Marcel Dekker, Inc.
- Pokhraj D., A. Lazarevic, R. L. Hoskinson, Z. Obradovic (2002). Spatial-Temporal Techniques for prediction and compression of Soil Fertility data. In: *Proceedings of the 6th International Conference on Precision Agriculture and Other Precision Resources Management*. Minneapolis, MN, USA .
- Pokhraj D., Z. Obradovic (2001). Improved Spatial-Temporal Forecasting through Modeling of Spatial Residuals in Recent History. In: *Proceedings First SIAM International Conference on data mining*. Chicago, USA .

Statistical models for retinal image matching

P. Puig¹, N. Adell¹, A. Rojas-Olivares², G. Caja², S. Carné²
and A.A.K. Salama²

¹ Servei d'Estadística, Universitat Autònoma de Barcelona

² Departament de Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona

Abstract: We present an example of random effects data analysis, related with a problem of animal identification, where the response is restricted to the interval (0,1) and data also present one-inflation. The models are based on Beta distributions.

Keywords: Retinal image recognition; Logistic-Normal distribution; One-inflated Beta distribution; Beta regression.

1 Introduction

Animal identification is a major requirement for government authorities and is a vital tool in tracing diseases of public and animal health. For instance, many countries have adopted national databases to monitor cattle movements.

Two biometric technologies, which can produce secure, unique identifiers are DNA profiling and retinal imaging. Nowadays retinal imaging is more feasible than DNA profiling. The retinal imaging recognition (RIR) is based on the uniqueness and invariability of the retinal vascular pattern of the eye.

With the aim of auditing the traceability of sheep, retinal image recognition (RIR) was used in a total of 152 lambs of 2 breeds (Lacaune, $n = 70$; Manchega, $n = 82$). Lambs wore ear tags and electronic boluses as controls. Retinal images were recorded twice from the 2 eyes of each lamb using an OptiReader device (Optibrand, Fort Collins, CO), a commercially available video camera expressly designed for capturing retinal vascular patterns of livestock. After a training period, the images were taken at 3 ($n = 152$), 6 ($n = 58$) and 12 month of age ($n = 58$). Digitalized images were treated by using the Optibrand Data Management Software (v. 4.1.3) and the 3 month enrolment images were used as the reference for further analysis. The response variable is the Optibrand's matching score (MS) percentage, a value between zero and one hundred (between zero and one, for us). The matching process uses an implemented matching algorithm depending on the degree of similarity in vessel size, vessel position and branch angles

observed between pairs of images. The higher the score, the more likely the images in the pair were from the same eye and from the same animal.

2 The One-inflated Beta distribution

First of all, we have analyzed the MS for the 3 month images in order to identify the distribution profile. There are many continuous distributions appropriate for analyze data lying in the interval $(0, 1)$ but the most popular are the Beta and the Logistic-Normal distributions.

However in this data set we have detected an excess of values equal to 1 and this does not correspond to the profile of a continuous distribution over $(0, 1)$. We suppose that this remarkable fact is due to the matching algorithm used by the Data Management Software (v. 4.1.3). We suspect that each image is partially analyzed at the beginning of the process and, if the matching is complete, the algorithm ends and it gives a MS value of 1. Otherwise, the algorithm continues analyzing the whole image and gives the corresponding MS value. Consequently, this is a quick algorithm that has been designed for perform about 20 pair matches per minute but provides MS values of 1 with a non zero probability.

Consequently, for analyze these kind of semicontinuous data, we have used the One-inflated Beta distribution. The One-Inflated Beta distribution (OIB), defined over $x \in [0, 1]$ has a pdf of the form,

$$f(x; \mu, \phi, p) = pI(x) + (1-p)(1-I(x)) \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} x^{\mu\phi-1} (1-x)^{(1-\mu)\phi-1}$$

Here $I(x)$ is an indicator function, that is $I(1) = 1$ and $I(x) = 0$ for $x \neq 1$. The right part of this pdf, that corresponds to the classical Beta distribution, has been parameterized according to Ferrari and Crivari-Neto (2004). The parameter μ is the population mean and ϕ is a precision parameter. In fact, this model is equivalent to the Zero-Inflated Beta distribution if each observation x is replaced by $1 - x$. The Zero-Inflated Beta models have been recently used for analyze proportions in finance (Cook et al., 2008). We also assume that for each observation different of one, the corresponding μ_i depends linearly on the covariates by means of an appropriate link function like logit, probit, etc. Moreover the one-inflation parameter p_i has also been modeled in terms of μ_i in order to reflect the empirical fact that high values of μ_i are accompanied by a high proportion of 1's. In addition, the empirical evidence of our analysis shows that the parameter ϕ is also related to μ_i . Consequently, after several trials we have concluded that the best model for our data set is the following:

$$x_i \sim OIB(\mu_i, \phi_i, p_i); \log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 z_i; \phi_i = \frac{c}{1-\mu_i}; p_i = \mu_i^\gamma$$

TABLE 1. Parameter estimates for the models of retinal image matching (age=3 month). The values in brackets are the standard errors.

Eye	Parameters			
	β_0	β_1	c	γ
Left	2.466 (0.125)	-0.371 (0.150)	0.994 (0.125)	11.114 (1.591)
Right	2.360 (0.141)	-0.212 (0.169)	0.929 (0.130)	7.589 (1.207)

Here z_i is a dichotomic variable that indicates the breed of the animal. For parameter estimation, we have maximized the corresponding log-likelihood function by using a program made in R that is available on request. The results given by this program are shown in Table 1. Notice that there is a slight significant effect of the breed (parameter β_1) for the left eye. It can be a spurious relation or it could also be due to an initial difficulty to handle the animals together with a systematic sampling following always the sequence left-right eye.

In the next section we are going to introduce parsimonious models in order to analyze together the matching scores coming from both eyes.

3 The One-inflated Bivariate Beta distribution

In order to study inter-eye images used as a traceability indicator, we have also considered random effects models. These models could be considered in a similar way to those described in Qiu et al. (2008), but with the special feature of the one-inflated terms. However, we have preferred to use a different way based on the internal structure of the beta distribution. It is well known that if U and W are independently gamma distributed random variables with the same scale parameter but different shape parameters, then $U/(U+W)$ follows a Beta distribution. Let (X, Y) be the two dimensional random vector that represents the MS observations from the left and right eyes of the same animal. We assume that this random vector can be written in the following way:

$$(X, Y) = \left(\frac{U}{U+W}, \frac{V}{V+W} \right),$$

where U , V and W are independently gamma distributed random variables with the same scale parameter and different shape parameters. Notice that U , V and W are not observable. The random variable W can be understood as the "animal" effect. Obviously, X and Y each follows a Beta distribution and they are correlated (see Olkin and Liu, 2003). The joint density, conveniently parameterized, is the following:

$$f(x, y; \mu_x, \mu_y, c) = \frac{x^{\frac{c\mu_x}{1-\mu_x}-1} y^{\frac{c\mu_y}{1-\mu_y}-1} (1-x)^{\frac{c}{1-\mu_y}-1} (1-y)^{\frac{c}{1-\mu_x}-1}}{B(\mu_x, \mu_y, c)(1-xy)^{c(\frac{\mu_x}{1-\mu_x} + \frac{\mu_y}{1-\mu_y} + 1)}}$$

Here, $E(X) = \mu_x$, $E(Y) = \mu_y$. Due to the special structure of this bivariate density, the precision parameters for X and Y are connected with their corresponding population mean and the parameter c in the following way: $\phi_x = c/(1 - \mu_x)$, $\phi_y = c/(1 - \mu_y)$. This is compatible with the empirical evidence of the univariate analysis.

The one-inflation phenomenon is more complicate for these bivariate patterns, because the 1's can appear in only one component or in both. If $f(x; \mu_x, c)$ and $f(y; \mu_y, c)$ are the marginal densities of X and Y , the pdf of the one-inflated Bivariate Beta distribution (OIBB) can be written as follows:

$$g(x, y) = \begin{cases} p_{11} & x = 1, y = 1 \\ p_{x1}f(x; \mu_x, c) & x \neq 1, y = 1 \\ p_{1y}f(y; \mu_y, c) & x = 1, y \neq 1 \\ (1 - p_{11} - p_{x1} - p_{1y})f(x, y; \mu_x, \mu_y, c) & x \neq 1, y \neq 1 \end{cases}$$

Notice that p_{11} , p_{x1} and p_{1y} indicate the proportion of observations of the form $(1, 1)$, $(x, 1)$ and $(1, y)$ respectively.

After several analysis we have found that the best model for our specific data set has the following form:

$$(x_i, y_i) \sim OIBB(\mu_{xi}, \mu_{yi}, c, p_{11}, p_{x1i}, p_{1yi}); p_{x1i} = \mu_{xi}^\gamma, p_{y1i} = \mu_{yi}^\gamma$$

$$\log\left(\frac{\mu_{xi}}{1 - \mu_{xi}}\right) = \beta_{x0} + \beta_{x1}z_i, \log\left(\frac{\mu_{yi}}{1 - \mu_{yi}}\right) = \beta_{y0} + \beta_{y1}z_i$$

The log-likelihood function has been maximized by using a program made in R and the results are shown in Table 2. For this full model the maximum of the log-likelihood function has been 26.995. In order to detect if there is any significant effect due to the eye (left or right), we have also fitted a restricted model where $\beta_{0x} = \beta_{0y} = \beta_0$ and $\beta_{1x} = \beta_{1y} = \beta_1$. The results are shown in Table 3. Now the maximum of the log-likelihood function has been 24.102 and, consequently, the likelihood ratio test gives a p-value equal to 0.055. This p-value is slow and indicates a slight influence of the effect eye.

We have used similar techniques for study the traceability of inter-age images, by constructing longitudinal models based on multivariate beta distributions. The final conclusion of our analysis is that retinal imaging recognition is a useful technology for auditing the identity of living lambs.

TABLE 2. Parameter estimates for the bivariate model of retinal image matching (age=3 month). The values in brackets are the standard errors.

Parameters						
β_{0x}	β_{1x}	β_{0y}	β_{1y}	γ	c	p_{11}
1.806	-0.166	2.006	-0.190	12.224	0.534	0.251
(0.136)	(0.137)	(0.134)	(0.129)	(1.449)	(0.053)	(0.035)

TABLE 3. Parameter estimates for the restricted bivariate model of retinal image matching (age=3 month). The values in brackets are the standard errors.

Parameters				
β_0	β_1	γ	c	p_{11}
1.889	-0.176	11.910	0.533	0.250
(0.122)	(0.107)	(1.406)	(0.053)	(0.035)

Acknowledgments: This research was partially supported by grant MTM2006-01477 from the Ministry of Education of Spain.

References

- Cook, D.O., Kieschnick, R. and McCullough, B.D. (2008). Regression analysis of proportions in finance with self selection *Journal of Empirical Finance*, **15**, pp. 860–867.
- Ferrari, S. L. P.; Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *J. Appl. Stat.*, **31**, no. 7, pp. 799–815.
- Olkin, I. and Liu, R. (2003). A bivariate beta distribution. *Statistics and Probability Letters*, **62**, pp. 407–412.
- Qiu, Z., Song, P.X.K. and Tan, M. (2008). Simplex Mixed-Effects Models for Longitudinal Proportional Data. *Scandinavian Journal of Statistics*, **35**, no. 4, pp. 577–596.

Instrumental Variable Estimation for Generalized Additive Models

Rosalba Radice¹ and Giampiero Marra²

¹ Corresponding author: Mathematical Sciences, University of Bath, Bath BA2 7AY, U.K. Email: r.radice@bath.ac.uk

² Mathematical Sciences, University of Bath, Bath BA2 7AY, U.K.

Abstract: Regression model literature has generally assumed that observable and unobservable covariates are statistically independent. However, for many applications this assumption is clearly tenuous. When unobservables are correlated with included regressors, standard estimation methods will not be valid. This means that estimation results from observational studies where unmeasured confounding is suspected to be present will be biased and inconsistent. One method for obtaining consistent estimates of treatment effects when dealing with linear models is the instrumental variable (IV) approach. However, linear models have been extended to generalized linear models (GLMs) and generalized additive models (GAMs), and although IV methods have been proposed to deal with GLMs, IV analysis has not been generalized to the GAM context. We propose a simple two-stage strategy for consistent IV estimation when dealing with GAMs represented using any penalized regression spline approach. We illustrate its empirical validity through an extensive simulation experiment and an air pollution study.

Keywords: Generalized additive models; Instrumental variables; Two-stage estimation procedure; Unmeasured confounding.

1 Introduction

Observational data are often used in statistical analysis to infer the effects of predictors of interest (treatments) on a particular response variable. The main characteristic of observational studies is a lack of treatment randomization which usually leads to selection bias. The most common solution to this problem is to account for confounding variables that are correlated with both treatment and response. However, the researcher might fail to adjust for pertinent confounders as they might be either unknown or not readily quantifiable. This constitutes a serious limitation to covariate adjustment since the use of standard estimators may yield biased and inconsistent estimates. Hence, a big concern when estimating treatment effects is how to account for unmeasured confounders.

This problem is known in applied economics as *endogeneity* of explanatory variables. The most commonly used econometric method to model data

that are affected by the unobservable confounding issue is the instrumental variable (IV) approach (Wooldridge, 2002). This method can yield consistent parameter estimates, but relies on the existence of one or more IVs that induce substantial variation in the endogenous treatment variable, have no direct effect on the response, and are uncorrelated with the unobservable confounders.

The applied and theoretical literature on the use of IVs in parametric and nonparametric regression models with Gaussian response is large and well understood. In many applications, however, Gaussian regression models have been replaced by GLMs and GAMs. Some solutions have been proposed to deal with GLMs in which selection bias is suspected. For example, the generalized method of moments and simultaneous maximum-likelihood estimation can be employed (Wooldridge, 2002).

IV analysis has not been extended to the GAM context and it is not clear how this can be achieved taking one of the approaches above. Here we extend the IV approach to GAMs by exploiting the two-stage approach proposed by Hausman (1978).

2 IV estimation for GAMs

A GAM can be written as:

$$\mathbf{y} = g^{-1}(\boldsymbol{\eta}) + \boldsymbol{\epsilon}, \quad \mathbb{E}(\boldsymbol{\epsilon}|\mathbf{X}) = \mathbf{0},$$

where $g(\cdot)$ is a link function, $g^{-1}(\boldsymbol{\eta}) = \boldsymbol{\mu}$, $\boldsymbol{\mu} \equiv \mathbb{E}(\mathbf{y}|\mathbf{X})$, \mathbf{y} is a vector of independent response variables following some exponential family distribution, $\boldsymbol{\eta}$ is the linear predictor, and $\boldsymbol{\epsilon}$ is an additive, unobservable error defined as $\boldsymbol{\epsilon} \equiv \mathbf{y} - g^{-1}(\boldsymbol{\eta})$. The linear predictor of a GAM is usually given by:

$$\boldsymbol{\eta} = \mathbf{X}^* \boldsymbol{\beta} + \sum_j \mathbf{f}_j(\mathbf{x}_j^+),$$

where $*$ and $+$ indicate discrete and continuous predictors and the \mathbf{f}_j are smooth functions of the covariates, \mathbf{x}_j^+ , represented using regression splines and subject to identifiability constraints. Here, $\mathbf{X}^* = (\mathbf{X}_e^*, \mathbf{X}_o^*, \mathbf{X}_u^*)$ where \mathbf{X}_e^* is a matrix of endogenous variables, \mathbf{X}_o^* a matrix of observable confounders, and \mathbf{X}_u^* a matrix of unobservable confounders that influence the response variable and are correlated with the endogenous predictors. $\boldsymbol{\beta}^\top = (\boldsymbol{\beta}_e^\top, \boldsymbol{\beta}_o^\top, \boldsymbol{\beta}_u^\top)$. Similarly, $\mathbf{X}^+ = (\mathbf{X}_e^+, \mathbf{X}_o^+, \mathbf{X}_u^+)$.

We can not observe \mathbf{X}_u^* and \mathbf{X}_u^+ and this violates the assumption that the error term is uncorrelated with the regressors, therefore leading to biased and inconsistent estimates. To this end, the endogenous variables have to be modelled:

$$\mathbf{x}_{ep} = g_p^{-1}\{\mathbf{Z}_p^* \boldsymbol{\alpha}_p + \sum_j \mathbf{f}_j(\mathbf{z}_{jp}^+)\} + \boldsymbol{\xi}_{up}, \quad p = 1, \dots, \quad (1)$$

where \mathbf{x}_{ep} represents the generic endogenous variable, $\mathbf{Z}_p^* = (\mathbf{X}_o^*, \mathbf{X}_{IVp}^*)$ with parameter vector $\boldsymbol{\alpha}_p$, $\mathbf{Z}_p^+ = (\mathbf{X}_o^+, \mathbf{X}_{IVp}^+)$, and $\boldsymbol{\xi}_{up}$ is a term containing information about structured and unstructured terms. Provided the instruments contained in \mathbf{X}_{IVp}^* and \mathbf{X}_{IVp}^+ meet the IV conditions, the $\boldsymbol{\xi}_{up}$ do contain information about the unmeasured confounders.

2.1 The two-step GAM estimator

We propose a Hausman-like approach. Two-step generalized additive model (2SGAM) procedure:

1. Fit (1) and then calculate:

$$\widehat{\boldsymbol{\xi}}_{up} = \mathbf{x}_{ep} - g_p^{-1} \{ \mathbf{Z}_p^* \widehat{\boldsymbol{\alpha}}_p + \sum_j \widehat{\mathbf{f}}_j(\mathbf{z}_{jp}^+) \}, \quad p = 1, \dots,$$

2. Fit a GAM defined by:

$$\mathbf{y} = g^{-1} \{ \mathbf{X}_{eo}^* \boldsymbol{\beta}_{eo} + \sum_j \mathbf{f}_j(\mathbf{x}_{jeo}^+) + \sum_p \mathbf{f}_p(\widehat{\boldsymbol{\xi}}_{up}) \} + \boldsymbol{\varsigma},$$

where $\mathbf{X}_{eo}^* = (\mathbf{X}_e^*, \mathbf{X}_o^*)$ and $\mathbf{X}_{eo}^+ = (\mathbf{X}_e^+, \mathbf{X}_o^+)$.

The 2SGAM estimator can be employed using GAMs represented using any penalized regression spline approach. We need to include the $\mathbf{f}_p(\widehat{\boldsymbol{\xi}}_{up})$ since we are interested in capturing the non-linear effects that unmeasured confounders have on the response. This would clear up the endogeneity of the $\mathbf{f}_j(\mathbf{x}_{jeo}^+)$.

This procedure yields consistent estimates, and in order to show this point it is convenient to recall that the smooth terms can be represented using regression spline type bases. In fact, the two steps can be written as:

1. Fit (1) and then calculate:

$$\widehat{\boldsymbol{\xi}}_{up} = \mathbf{x}_{ep} - g_p^{-1} (\mathbf{Z}_p \widehat{\boldsymbol{\delta}}_p), \quad p = 1, \dots, \quad (2)$$

where $\mathbf{Z}_p \widehat{\boldsymbol{\delta}}_p = \mathbf{Z}_p^* \widehat{\boldsymbol{\alpha}}_p + \sum_j \widehat{\mathbf{f}}_j(\mathbf{z}_{jp}^+)$.

2. Fit a GAM defined by:

$$\mathbf{y} = g^{-1} (\mathbf{X}_e \boldsymbol{\beta}_e + \mathbf{X}_o \boldsymbol{\beta}_o + \widehat{\boldsymbol{\Xi}}_u \boldsymbol{\beta}_{\Xi_u}) + \boldsymbol{\varsigma}, \quad (3)$$

where $\mathbf{X}_e \boldsymbol{\beta}_e + \mathbf{X}_o \boldsymbol{\beta}_o = \mathbf{X}_{eo}^* \boldsymbol{\beta}_{eo} + \sum_j \mathbf{f}_j(\mathbf{x}_{jeo}^+)$ and $\widehat{\boldsymbol{\Xi}}_u \boldsymbol{\beta}_{\Xi_u} = \sum_p \mathbf{f}_p(\widehat{\boldsymbol{\xi}}_{up})$.

Now, supposing the $\boldsymbol{\delta}_p$ were known, then by using (1) the column vectors of $\boldsymbol{\Xi}_u$ would be known. Hence the information about the unobservables could be incorporated into the model by considering $\boldsymbol{\Xi}_u$. In this respect,

the endogeneity issue would disappear since the assumption that the error term is uncorrelated with the predictors would be satisfied. However, we do not know the δ_p . By using (2) we can get consistent estimates for the δ_p thereby obtaining a good estimate for Ξ_u . It can be readily shown that $\widehat{\beta}^\top$, now defined as $(\widehat{\beta}_e^\top, \widehat{\beta}_o^\top, \widehat{\beta}_{\Xi_u}^\top)$, is consistent for the vector value $\gamma^\top = (\gamma_e^\top, \gamma_o^\top, \gamma_u^\top)$ that optimizes:

$$\mathbb{E}[\{\mathbf{y} - g^{-1}(\mathbf{X}_e\gamma_e + \mathbf{X}_o\gamma_o + \Xi_u\gamma_u)\}^2]. \quad (4)$$

In (4) we ignore estimation for Ξ_u as the endogeneity issue only concerns the second-step equation, and because consistent estimates for it can be obtained. All we need to show is that β^\top is equal to the value of γ^\top , and provided the IVs meet the assumptions discussed in Section 1, we have that:

$$\mathbb{E}(\mathbf{y}|\mathbf{X}_e, \mathbf{X}_o, \mathbf{X}_{IV}) = g^{-1}(\mathbf{X}_e\beta_e + \mathbf{X}_o\beta_o + \Xi_u\beta_{\Xi_u}).$$

It follows that $\beta = \gamma$, hence β is also the optimizer of (4). The formal proof of the arguments above (omitted here) consists of adapting the standard results for two-step non-linear estimators. These can be found in Wooldridge (2002).

3 Results

The empirical properties of 2SGAM are shown through a Monte Carlo simulation study. Also, an application to air pollution data, which illustrates our approach, is briefly discussed.

3.1 Simulation results

Figures 1, 2 and 3 show some of the simulation results. Generally speaking, 2SGAM performs well for finite sample sizes, and as the number of observation increases the estimates converge to the true population curves. 2SGAM performs well as long as the IV is strong enough. In fact, as pointed out by Bound, Jaeger and Baker (1995), IV methods can be ill-behaved if the instruments are not highly correlated with the endogenous variables of interest (see Figure 3).

3.2 Empirical results

We applied the procedure to the National Morbidity, Mortality, and Air Pollution Study database for the city of Pittsburgh to estimate the association between air pollution and mortality, in the presence of confounding factors which can not be controlled for using a smooth function of time.

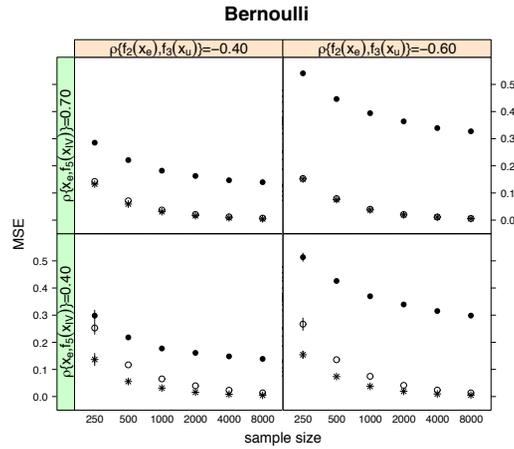


FIGURE 1. Mean squared error (MSE) results for $\hat{f}_2(x_e)$ (see Figure 2) when data are simulated from a Bernoulli distribution. \circ indicates the 2SGAM estimator results, whereas \bullet and $*$ refer to the cases in which estimation is carried out without accounting for unmeasured confounding (naive GAM), and that in which the unobservable variable is available and used during the model-fitting process. Penalized thin plate regression splines were employed (Wood, 2006). The vertical lines show ± 2 standard error bands, which are only reported for the cases in which they are substantial. Notice the good overall performance of the proposed method for all sets of correlations and sample sizes.

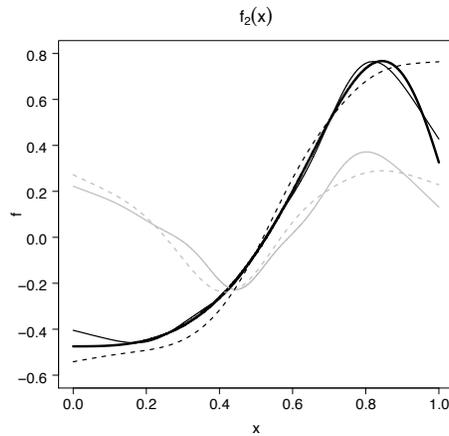


FIGURE 2. Typical estimated smooth functions for $f_2(x_e)$ (thicker solid black line) when employing the 2SGAM procedure (black lines) and naive GAM estimation (grey lines). The dotted lines indicate the results for the cases in which $n = 1000$, whereas the solid lines refer to the cases in which $n = 8000$. Notice the convergence of the proposed method to the true function as opposed to the naive approach.

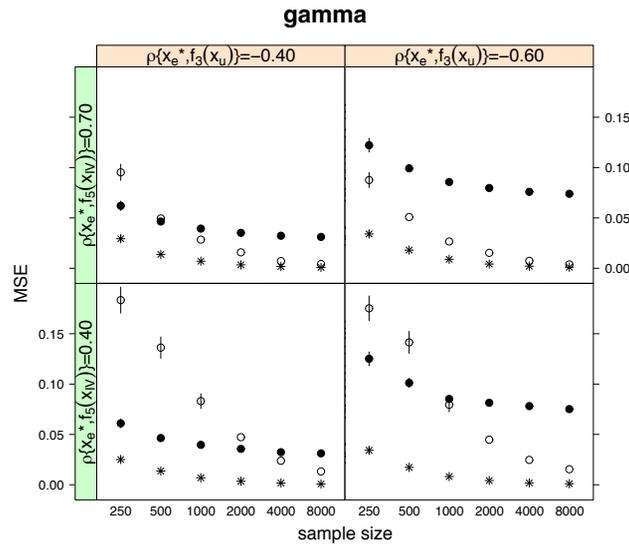


FIGURE 3. MSE results for $\hat{\beta}_e$ when data are simulated from a gamma distribution. Here x_e is a binary variable, hence more information is needed to get the parameter estimate right. Notice that for low sample sizes the naive method seems to outperform 2SGAM when the instrument is not strong.

Our results provide evidence in support of the fact that the air pollution-mortality relationship is not robust to the presence of unmeasured confounders which vary on timescales which are different from those of pollution and mortality. Hence important questions remain as to whether this relationship should be reanalyzed for many other cities using the approach proposed in this article.

References

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, **90**, 443–450.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, **46**, 1251–1271.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. London: Chapman & Hall.

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

Improved SNP genotyping using model-based calibration

Ralph C.A. Rippe¹, Paul H.C. Eilers², Pim J. French³ and
Jacqueline J. Meulman⁴

¹ Department 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

³ Department of Neurology, Erasmus Medical Center, Rotterdam, The Netherlands

⁴ Department of Mathematics, Leiden University, Leiden, The Netherlands

Abstract: SNP arrays are used to determine DNA composition (genotyping), using fluorescence data to measure concentrations. Strong systematic patterns occur in the SNP signals. We quantify these patterns and use them to improve genotyping and show some more applications to real tumor samples. Since we have a large-scale problem at hand, we use specialized algorithms.

Keywords: DNA; Fluorescence; Mixtures; Calibration.

1 Introduction

Single nucleotide polymorphisms (SNPs, pronounced as “snips”) are mutations in which only one of the bases (A, C, G or T) that make up DNA has changed. Human DNA chromosomes are present in pairs. Each SNP is therefore present on two alleles which results in an AA, AB or BB genotype. SNPs are measured by hybridization (attachment) of DNA fragments from biological samples to several “probes” for a single SNP. The strength of the fluorescence signal indicates the type of hybridized DNA fragments. After some processing, we can get two signals, one for each strand. If we combine these signals, we can apply the following theorem: in general, a signal with strength a (b) is observed when allele A (B) is present. So, when the genotype is AA (BB), no signal b should be observed, but only signal a of double strength. When the genotype is AB, we should observe both a and b , with equal strengths. The goal of the measurements is to reliably determine genotypes. Present (commercial) SNP analysis programs have a high noise ratio and (apart from SNP6.0) do not take the samples allelotype. Improvement of both would result in a better definition of the chromosomal aberrations and, in cancer, a better definition of candidate

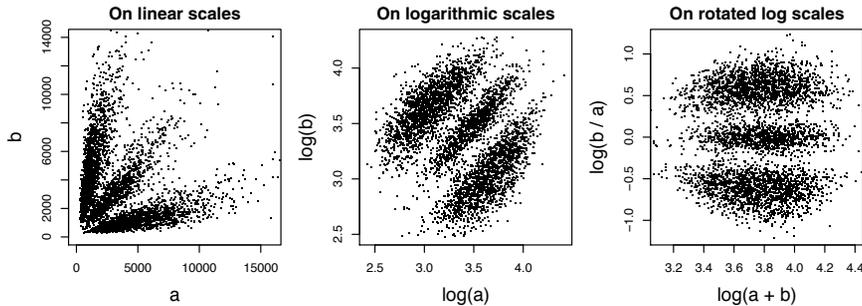


FIGURE 1. Data (Affymetrix 50k Hind, available from HapMap) transformation for signals on chromosome 9. Signal a (b) represents allele A (B).

oncogenes and tumoursuppressor genes that may reside on the chromosomal region. The scale of the data can be enormous: from 1500 to 10^6 SNP measurements on up to many thousands of samples. In the following text, we use 50k Affymetrix samples from the HapMap repository and 250k Affymetrix arrays collected in the Erasmus Medical Center in Rotterdam, The Netherlands.

From (a, b) we compute a transformed pair of variables (u, v) (Figure 1). In the plane of u and v the three genotypes form three clearly visible clusters. This allows us to use a mixture of three regression lines as a model for the data. After fitting the model, membership probabilities indicate the most probable genotype for each SNP. There is some overlap between clusters. We have found that remarkably stable systematic fluorescence patterns occur, and especially that the relative level of fluorescence of each SNP, corrected for genotype, shows little variation. The patterns can be described faithfully by a linear model with three sets of parameters, one for the SNPs, another for the biological samples and a third for the genotypes.

2 Models and calibration

2.1 Global and local model

We transform our data to $u = \log(a+b)$ and $v = \log(a/b)$ (logs to base 10). As the right panel of Figure 1 shows, three clusters are present, and each cluster can be approximated well by a straight line or a quadratic curve surrounded by noise. The shape of the clusters in the raw signals depends on the technology used; for Affymetrix signals, the clusters are usually quadratic, whereas raw signals from e.g. Illumina or corrected signals, independent of the platform used, show linear relationships. The R package *FlexMix* makes it very easy to estimate the mixture of regression lines. This is shown in the left (quadratic) and right (linear) panels in Figure 2. The

dots have been colored by mixing the colors red, blue and green proportionally to the membership probabilities. Most SNPs are classified well, but quite a number of them fall in the area between the clusters. We will exploit systematic patterns in fluorescence strengths to get better separation.

To develop the model we assume that genotypes are known. Let $x_{ij} = \log(a_{ij})$, with a_{ij} the fluorescence for allele A of SNP i on array (= biological sample) j , where $i = 1, \dots, m$, $j = 1, \dots, n$, m is the number of SNPs and n the number of arrays. Let the genotypes be coded as the indicator array $H = [h_{ijk}]$. Our first model is

$$\hat{x}_{ij} = \mu + \alpha_i + \beta_j + \sum_{k=1}^3 \gamma_k h_{ijk}, \quad (1)$$

where μ is the grand mean, α_i describes the overall level of SNP i , β_j describes the overall level of sample j and γ_{kc} is characteristic for genotype k . We introduce constraints $\sum_i \alpha_i = 0$ and $\sum_j \beta_j = 0$. We call this the global model (Rippe et al., 2009a), since it has one set of genotype parameters (γ) for all SNPs. A refinement is to have separate genotype parameters for each SNP. We call this the local model (Rippe et al., 2009b), which is specified as

$$\hat{x}_{ij} = \mu + \beta_j + \sum_{k=1}^3 \gamma_{ik} h_{ijk}. \quad (2)$$

We constrain $\sum_j \beta_j = 0$. Equivalent models are specified for the B allele, with $y_{ij} = \log(b_{ij})$.

2.2 Parameter estimation

The type of array we are analyzing here covers up to a quarter of a million SNPs. Datasets contain from 10 to 100s of arrays. So we have millions of datapoints and a huge number of parameters: approximately 250.000 in the global model and three times more in the local model. Our models can be written as regression models, but explicit construction of the design matrix and invoking a regression procedure is not a good idea: the design matrix would have 10^9 to 10^{13} elements. The design matrix is very sparse, so the next best solution is to use sparse matrix software. We have not tried this approach, so we cannot report on its effectiveness. Instead we have explored block relaxation and symbolic solutions of the normal equations that follow from linear regression.

In both models (1) and (2) it is easy to compute a set of parameters once the rest is available. One simply has to average residuals, over SNPs, arrays or genotypes, dependent on the type of parameters. Starting from reasonable starting values (averages over SNPs for α , averages over arrays for β), one iteratively updates each set of parameters. In the numerical analysis literature this is known as block relaxation. In our experience the speed

of convergence is quite good: from 10 to 30 iterations generally suffice to make the updates of the order of 10^{-6} (relative size). The constraints on α and β are applied in each iteration.

It is not hard to build and solve the normal equation symbolically. We illustrate this for the local model. With appropriate B and C , we can write

$$y = B\beta + C\gamma + e \tag{3}$$

where β contains the β parameter in (2) and $\gamma = \text{vec}(\Gamma)$, i.e. the columns of $\Gamma = [\gamma_{ik}]$ below each other, and $y = \text{vec}(Y)$. The structure of B is simple, it can be written as $B = I_n \otimes 1_m$, where I_n is the n -by- n identity matrix and 1_m is a vector of ones, of length m . The structure of C is more complex, it consists of n rows and three columns of diagonal matrices. If we indicate these by C_{jk} , then C_{jk} has on its diagonal column j in layer k of the indicator array H .

We do not form B and C explicitly. Instead we study the normal equations

$$\begin{bmatrix} B'B & B'C \\ C'B & C'C \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} B'y \\ C'y \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} \tag{4}$$

We can prove that $B'B = mI_n$, $C' = \tilde{H}$, $C'C = D$, where \tilde{H} is a matrix formed by placing the three layers of H below each other. D is a $3m$ by $3m$ diagonal matrix; its first (second, third) m diagonal elements gives, for each SNP, the number of times genotype 1 (2,3) occurs. Furthermore we have that $B'y$ contains the sums of the columns of Y , while $C'y$ is a stack of three vectors; the first (second, third) vectors contain the sum, per SNP of the elements of y corresponding to genotype 1 (2,3).

From (4) follows: $\hat{\gamma} = V_{22}^{-1}(r_2 - V_{21}\hat{\beta})$ and hence $(V_{11} - V_{12}V_{22}^{-1}V_{21})\hat{\beta} = r_1 - V_{12}V_{22}^{-1}r_2$. Because V_{22} is a diagonal matrix, multiplication by V_{22}^{-1} boils down to dividing the rows of a vector or matrix by the corresponding diagonal elements of V_{22} . Hence, it is light work to compute $V_{11} - V_{12}V_{22}^{-1}V_{21}$ and solve for $\hat{\beta}$, a vector of moderate length. Additional efficiency can be realized by exploiting the way V_{21} is formed. We omit the details here.

In this analysis we have ignored the fact that in the system in (4) is singular, because the condition $\sum_j \beta_j = 0$ is not applied. An easy way out is to demand the minimum-norm solution for β , by replacing $B'B$ in (4) by $B'B + \kappa I$ with κ a small number.

2.3 Calibration and classification

From the model follows, for each allele, either a parameter vector $\alpha = [\alpha_i]$ or a three-column array $\Gamma = [\gamma_{ij}]$. We can use them to compute corrected signals and repeat the estimation of the mixture model. We call this calibration. In the first case this translates to $a_{ij}^* = a_{ij}/10^{\alpha_i}$. It does not use the genotypes of a new sample, so we call it genotype-free calibration. In

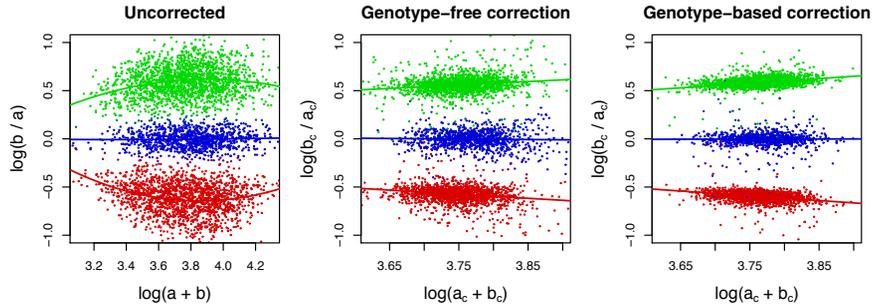


FIGURE 2. Correction effects for an example Affymetrix (HapMap) 50k Hind array. The left panel shows the uncorrected clusters for chromosome 9. The middle panel shows the clusters after genotype-free correction. The right panel shows the clusters after one round of genotypebased correction. Note the changed range on the x-axis in the middle and right panel.

the second case we have $a_{ij}^* = a_{ij}/10^\phi$, where $\phi = \sum_k h_{ijk}\gamma_{ik}$, which we call genotype-based calibration.

The latter approach selects the right calibration value based on the genotype at hand in a particular SNP. Since samples can have different genotypes, this method is also sample-specific. Analogous formulas hold for signal b . We propose to estimate either α or Γ for a set of high-quality samples. With the so-obtained α signals on a new array can be calibrated genotype-free.

The genotype is determined by the cluster that gives the highest membership probability for a SNP. Classification performance is measured via the proportion of membership probabilities < 0.80 , per sample, indicated by $Q80$. $Q80$ gives us the proportion of data points that were classified with relatively large uncertainty within a sample. Alternatively, if a stricter criterion is preferred, $Q90$ and $Q95$ can be used. However, all three proposals can identify the same badly fitted samples, if any occur at all. Therefore, using just one of these criteria is sufficient for initial evaluation.

3 Results

Table 1 summarizes the results for several 50k Affymetrix Hind arrays after applying both calibration methods. Figure 2 provides an illustration for one of these arrays.

Based on the global model we see a effective correction. It can be applied very generally, since it does not require any known genotypes. Furthermore, it is an important method when working with tumor data. The genotype-based calibration results in a even better cluster distinction. However, this

TABLE 1. Q-bands (proportion of cluster membership probabilities $< Q$, with $Q = \{0.80, 0.90, 0.95\}$) for several Affymetrix 50k Hind arrays. The Q80, Q90 and Q95 criteria drop at different levels for the two correction methods. The genotype-based correction gives the best overall results.

	Q80	Q90	Q95
Uncorrected data	6.0	8.6	12.3
Genotype-free correction	1.1	2.2	5.3
Genotype-based correction	0.2	0.4	0.8

approach requires an initial genotype classification, making it less generic. In the current stage of our work, it only work in diploid situations.

In general we can say that compared to the uncorrected data, both methods result in improved cluster separation.

The idea of improvement using a two-step procedure (genotype-free followed by genotype-based) or an iterative approach seems very attractive. So after correction with the global parameters, the mixture can be estimated and then Γ is used for further calibration, until convergence. Furthermore, this approach can also be used in signals from tumors, where the genotype can be complex.

4 Discussion

Bengtsson *et al* (2008), Giannoulatou *et al* (2008) and Xiao *et al* (2007) propose elaborate and time-consuming methods to improve genotyping (with or without data correction), where our method is simple, fast and effective. Further, all preliminary tests show that our approach doesn't depend on the platform used for SNP analysis. We need to look into the two-step or iterative approach in more detail.

In addition to improved cluster separation using both correction methods, we can also use the α parameters obtained from the reference samples to gain insight into so-called Copy Number Variations (CNV) and Loss of Heterozygosity (LOH) as described in e.g. Zhao *et al* (2004).

We assumed that the two alleles are present (diploidy) as is usually the case for healthy tissue and blood. However, during tumor growth, one or both alleles can be deleted, erroneously copied or mutated. In case of deletion this can result in LOH, where the missing allele is copied from the remaining allele, if present. We may end up with just a single chromosome or, in the worst case, we can even have (parts of) a chromosome completely missing. On the other hand, a combination of more than two alleles can also occur, e.g. AAB, BAB or even ABBB. These effects in relation to their chromosomal position (which translates to the physical location of genes on a chromosome), can be illustrated much clearer after correction and can be

referenced against crosstested data, like those from McCarroll *et al* (2006). Looking at the $\log(a + b)$ signal (Figure 3) and the $\log(b/a)$ signal (Figure 4) after correction separately, both LOH and CNV are clearly different from the normal signal. The correction is more effective when the signal is stronger, since a stronger signal has less noise. Figure 4 shows the effect on a sample from the Erasmus Medical Center and a control sample from Affymetrix, which has a higher quality than the tumor sample.

In our future research we aim to use a “Golden Standard” from the HapMap project to evaluate classification performance on high-quality data.

We will also investigate to what extent initial misclassification can persist, because in principle a point can be calibrated (using the genotype-based method) into the “wrong” cluster due to misclassification.

References

- Bengtsson, H., Irizarry, R., Carvalho, B. and Speed, T.P. (2008). Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics*, **24(6)**, 759-767.
- Giannoulatou, E., Yau, C., Colella, S., Ragoussis, J., & Holmes, C. (2008). Genosnp: a variational bayes within-sample snp genotyping algorithm that does not require a reference population. *Bioinformatics*, **24(19)**, 2209-2214.
- McCarroll S.A., et al. (2006). Common deletion polymorphisms in the human genome. *Nat. Genet.*, **38**, 86-92.
- Rippe, R.C.A., Eilers, P.H.C. & Meulman, J.J. (2008). Models for Fluorescence Signals on SNP Arrays. *Proceedings of the 23d International Workshop on Statistical Modelling*, 370-375.
- Rippe, R.C.A., Eilers, P.H.C. & Meulman, J.J. (2009a). Psychometric Modeling of Structure in Fluorescence Intensities of SNP Arrays. *submitted*.
- Rippe, R.C.A., Eilers, P.H.C. & Meulman, J.J. (2009b). SNP calibration on Illumina BeadArrays. *submitted*.
- Xiao, Y., Segal, M.R., Yang, Y.H., & Yeh, R.F. (2007). A multi-array multi-snp genotyping algorithm for affymetrix snp microarrays. *Bioinformatics*, **23-12**, 1459-1467.
- Zhao, X., et al. (2004). An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res.*, **65**, 5561-5570.

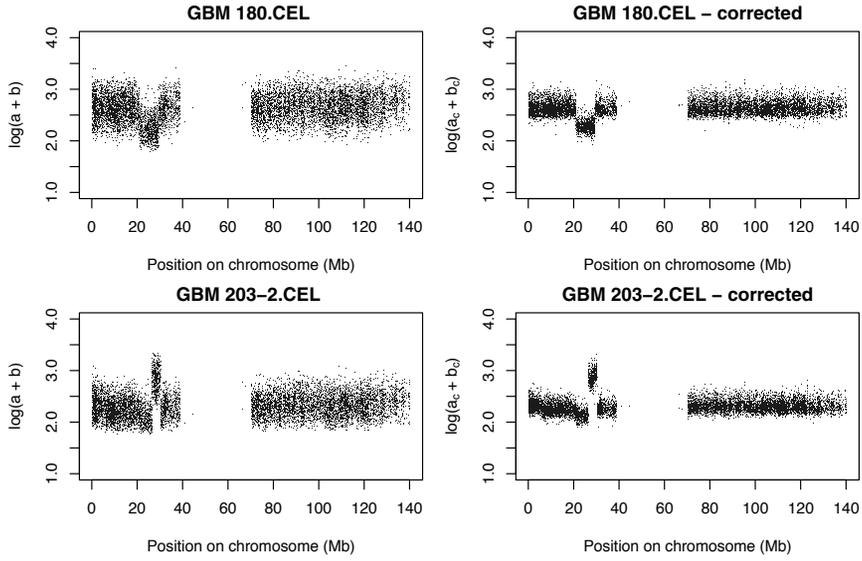


FIGURE 3. Data on chromosome 9. Illustration of CNV: the difference in amplitude indicates either a decrease in number of alleles (top) or an increase (bottom).

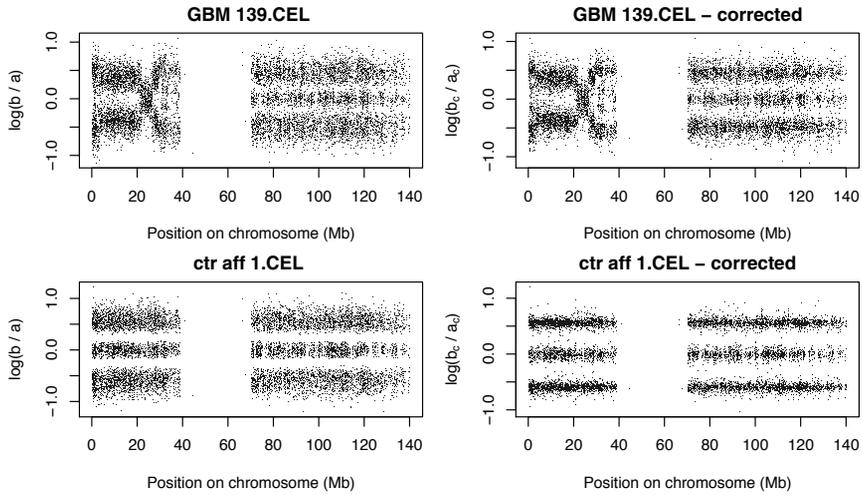


FIGURE 4. Data on chromosome 9. Illustration of LOH. In the top panels we see just two bands in the left side signal, whereas in the bottom panels we see three bands (one for each genotype) in a control sample. Deviations to the three-band signal indicate problematic DNA.

Modeling Time Varying Parameter Models Using Mixtures

Ori Rosen¹, Sally Wood² and Robert Kohn³

¹ Department of Mathematical Sciences, University of Texas at El Paso, El Paso, Texas 79968, U.S.A., E-mail: ori@math.utep.edu

² Melbourne Business School, University of Melbourne, Victoria, 3053, Australia

³ Robert Kohn, School of Economics and School of Banking and Finance, University of New South Wales, Sydney, New South Wales, 2052, Australia

Abstract: We develop a model for analyzing a time series whose parameters evolve over time or experience a small number of abrupt changes. We allow for parameter evolution using a mixture model whose components are time series with constant but unknown parameters and mixture probabilities that depend on time. The number of mixture components is determined from the data. We take a Bayesian approach which is implemented using Markov chain Monte Carlo sampling.

Keywords: Mixture model; Southern oscillation index; Time series; Reversible jump MCMC.

As a motivating example, consider the El Niño/Southern Oscillation (ENSO) phenomenon. ENSO is an irregular low frequency oscillation between a warm El Niño state and a cold La Niña state. The Southern Oscillation Index (SOI) is an indicator of the ENSO phenomenon and is calculated to be the standardized anomaly of the mean sea-level pressure difference between Tahiti and Darwin. The strong El Niños of 1982/83 together with the more frequent occurrences of El Niño in recent decades have raised the question of whether human-induced global warming has changed the structure of the ENSO time series. (Timmermann et al., 1999). The data, shown in Figure 1, are monthly values of the SOI from January 1876 to April 2008 and are available at <http://www.bom.gov.au/climate/current/soihtml.shtml>. There are several papers related to our methodology. A basic reference for regression mixture models having covariates in both the components and the component probabilities is Jacobs et al. (1991) who called them mixture-of-experts models. Rosen, Stoffer and Wood (2009) estimate an evolving spectral density by partitioning the data into segments of contiguous observations, calculating the log periodogram of each segment and then fitting a smoothly varying mixture to these log periodograms. Lao and So (2008) propose a Bayesian mixture of autoregressive models using a Dirichlet process priors. Unlike our approach, the mixing weights in their model are not a function of time.

Fitting piecewise autoregressive models was suggested by Kitagawa and Akaike (1978) who used AIC to determine the change points. Davis, Lee and Rodriguez-Yam (2006) proposed fitting piecewise autoregressive mod-

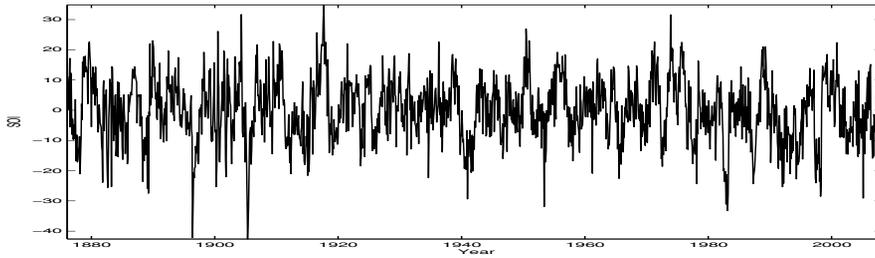


FIGURE 1. Monthly values of the Southern Oscillation Index (SOI) from January 1876 to April 2008

els using minimum description length and used genetic algorithms for solving the resulting optimization problem. Ombao et al. (2001) segment a time series using orthogonal complex-valued transforms that are simultaneously localized in time and space.

A second approach which allows the parameters to change is to model their evolution. For example, West et al. (1999) allow the parameters of an autoregressive process to change over time by modeling them as a random walk. However, they assume that the maximum lag in the autoregressive process is fixed. The assumption of a fixed lag is relaxed by Prado and Huerta (2002). Gerlach, Carter and Kohn (2000) provide a sampling scheme that allows for smooth parameter evolution as well as structural breaks in the parameters.

1 The model

To allow for parameter evolution and structural breaks for a time series $\{y_t, t = 1, \dots, n\}$ generated by a model having a predictive density $p(y_t|y_{t-1}, \dots, y_1; \omega)$, we propose the following mixture model, based on this predictive density

$$p(y_t|y_{t-1}, \dots, y_1; \theta_r) = \sum_{j=1}^r \pi_{tjr} p(y_t|y_{t-1}, \dots, y_1; \omega_{jr}) . \tag{1}$$

The mixing weights π_{tjr} depend on time and have the multinomial logit form

$$\pi_{tjr} = \frac{\exp(\delta'_{jr} \mathbf{s}_t)}{\sum_{h=1}^r \exp(\delta'_{hr} \mathbf{s}_t)} ,$$

where $\mathbf{s}_t = (1 \ t)'$, and for identifiability we take $\delta_{1r} = \mathbf{0}$. Let $\delta_r = (\delta'_{2r}, \dots, \delta'_{rr})$ and $\theta_r = (\delta'_{1r}, \omega'_{1r}, \dots, \omega'_{rr})'$. The number of components r is unknown but is determined from the data with an upper bound of R . Model (1) allows for possible structural breaks in the parameters or slowly

evolving parameters without having to directly model the change points or the evolution of the parameters.

We illustrate our approach using a mixture of autoregressive models. We write the j th component model as

$$y_t = \phi_{jr,0} + \sum_{k=1}^p \phi_{jr,k} y_{t-k} + \sigma_{jr} e_{jt}, \quad e_{jt} \sim N(0, 1).$$

For simplicity, we assume that all components have the same unknown lag length p . Thus, $\boldsymbol{\omega}_{jrp} = (\phi_{jr,0}, \phi_{jr,1}, \dots, \phi_{jr,p}, \sigma_{jrp}^2)'$, where the dependence on p is indicated by the additional subscript.

2 Prior specification

Prior on $\boldsymbol{\delta}_{rp} = (\boldsymbol{\delta}'_{2rp}, \dots, \boldsymbol{\delta}'_{rrp})'$:

The priors on $\boldsymbol{\delta}_{jrp}$, $j = 2, \dots, r$, are independent multivariate normal $N(\mathbf{0}, \sigma_\delta^2 I_2)$, where $\sigma_\delta^2 = n$.

Priors on r and p :

The maximum number of components is R and $\Pr(j = r) = 1/R$ for $r = 1, \dots, R$. Similarly, the maximum value of the lag is P and $\Pr(k = p) = 1/P$ for $p = 1, \dots, P$.

Priors on $\boldsymbol{\phi}_{rp} = (\boldsymbol{\phi}'_{1rp}, \dots, \boldsymbol{\phi}'_{rrp})'$:

The priors on the $\boldsymbol{\phi}_{jrp}$'s are Zellner's G-prior distributions (see Marin and Robert, 2007) $N(\mathbf{0}, c\sigma_{jrp}^2 (X_p' X_p)^{-1})$ where $c = n$ and

$$X_p = \begin{pmatrix} 1 & y_P & y_{P-1} & y_{P-2} & \dots & y_{P-p+1} \\ 1 & y_{P+1} & y_P & y_{P-1} & \dots & y_{P-p+2} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & y_{n-1} & y_{n-2} & y_{n-3} & \dots & y_{n-p} \end{pmatrix}.$$

Note that X_p is an $(n - P) \times (p + 1)$ matrix with a fixed number of rows and a variable number of columns, for $p = 1, \dots, P$.

Priors on $\boldsymbol{\sigma}_{rp}^2 = (\sigma_{1rp}^2, \dots, \sigma_{rrp}^2)'$:

The priors on the σ_{jrp}^2 's are independent inverse gamma distributions with densities $p(\sigma_{jrp}^2) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma_{jrp}^2)^{-(\alpha+1)} \exp(-\beta/\sigma_{jrp}^2)$, where $\alpha = \beta = 0.01$. This choice of α and β reflects vague knowledge of σ_{jrp}^2 . For identifiability, the σ_{jrp}^2 s are ordered.

3 Sampling scheme

We sample all the parameters, including r and p , from their posterior distribution, in two stages. In stage I, we run R separate chains for $r = 1, \dots, R$. In each of these chains, r is fixed while all other parameters, including p ,

are sampled. The results from this preliminary analysis are utilized in turn in stage II to perform a reversible jump step corresponding to varying values of r and p . To simplify the sampling scheme of Stage I, we introduce binary latent variables z_{sj} , such that $z_{sj} = 1$ if y_t is generated by the j th component, and $z_{sj} = 0$ otherwise.

3.1 Stage I: fixed r

1. Fix r and initialize \mathbf{z} , the vector of indicators.
2. Draw the lag p from the multinomial distribution $p(p|y^*, r, \mathbf{z})$, where $y^* = (y_{P+1}, \dots, y_n)'$.
3. For $j = 1, \dots, r$, draw σ_{jp}^2 from the inverse gamma distribution $p(\sigma_{jp}^2|y^*, \mathbf{z}_j, r, p)$.
4. For $j = 1, \dots, r$, draw ϕ_{jp} from the multivariate normal distribution $p(\phi_{jp}|\sigma_{jp}^2, \mathbf{z}_j, y^*, p)$.
5. For $j = 2, \dots, r$, draw δ_{jp} from the multivariate normal distribution $p(\delta_{jp}|\mathbf{z})$.
6. Draw \mathbf{z} from the multinomial distribution $p(\mathbf{z}|\phi_p, \sigma_p^2, \delta, r, p)$.

We run R chains for $r = 1, \dots, R$. From the iterates of each of these chains, we obtain the posterior distribution of the lag, as well as the posterior mean vectors $\hat{\phi}_{rp}$, $\hat{\delta}_{rp}$ and $\hat{\nu}_{rp} = \log \hat{\sigma}_{rp}^2$. The corresponding variance-covariance matrices $\hat{\Sigma}_{\phi_{rp}}$, $\hat{\Sigma}_{\delta_{rp}}$ and $\hat{\Sigma}_{\nu_{rp}}$ are obtained as sample variance-covariance matrices based on the iterates of ϕ_{rp} , δ_{rp} and $\nu_{rp} = \log \sigma_{rp}^2$, respectively.

3.2 Stage II: variable r and p

Stage II consists of a reversible jump step, corresponding to the values r and p of the unknown number of components and autoregressive lag. Specifically, the Metropolis-Hastings step is performed as follows.

1. Draw $r^{(n)}$ from a discrete uniform distribution over $\{1, 2, \dots, R\}$.
2. Draw $p^{(n)}$ from the posterior distribution of the lag over $\{1, 2, \dots, P\}$ obtained from Stage I.
3. Draw a vector $\phi^{(n)}$ from the multivariate normal distribution $N(\hat{\phi}_{r^{(n)}p^{(n)}}^{(n)}, \hat{\Sigma}_{\phi_{r^{(n)}p^{(n)}}^{(n)}})$, where $\hat{\phi}_{r^{(n)}p^{(n)}}^{(n)}$ and $\hat{\Sigma}_{\phi_{r^{(n)}p^{(n)}}^{(n)}}$ are from stage I.
4. Draw a vector $\delta^{(n)}$ from the multivariate normal distribution $N(\hat{\delta}_{r^{(n)}p^{(n)}}^{(n)}, \hat{\Sigma}_{\delta_{r^{(n)}p^{(n)}}^{(n)}})$, where $\hat{\delta}_{r^{(n)}p^{(n)}}^{(n)}$ and $\hat{\Sigma}_{\delta_{r^{(n)}p^{(n)}}^{(n)}}$ are from stage I.

5. Draw a vector $\boldsymbol{\nu}^{(n)}$ from the multivariate normal distribution $N(\hat{\boldsymbol{\nu}}_{r^{(n)}p^{(n)}}, \hat{\Sigma}_{\nu_{r^{(n)}p^{(n)}}})$, where $\hat{\boldsymbol{\nu}}_{r^{(n)}p^{(n)}}$ and $\hat{\Sigma}_{\nu_{r^{(n)}p^{(n)}}$ are from stage I.

4 Forecasting

Our goal is to improve prediction of future observations based on the mixture model (1), compared to prediction based on a model with a single component ($r = 1$). For a time series $\mathbf{y} = (y_1, \dots, y_n)'$ with y_t modeled by a density function $p(y_t|y_{t-1}, \dots, y_1; \boldsymbol{\omega})$ indexed by a parameter vector $\boldsymbol{\omega}$, the m -step-ahead predictive distribution is

$$p(y_{n+m}|\mathbf{y}) = \int p(y_{n+m}|\mathbf{y}, \boldsymbol{\omega})p(\boldsymbol{\omega}|\mathbf{y})d\boldsymbol{\omega},$$

where $p(\boldsymbol{\omega}|\mathbf{y})$ is the posterior distribution of $\boldsymbol{\omega}$. For an autoregressive model of order p , the density $p(y_{n+m}|\mathbf{y}, \boldsymbol{\omega})$ is normal with mean and variance μ_m and σ_m^2 , respectively, which can be computed via the Kalman filter. Using the MCMC iterates $(r^{(l)}, p^{(l)}, \boldsymbol{\delta}_{rp}^{(l)}, \boldsymbol{\omega}_{rp}^{(l)})$, an m -step-ahead prediction based on our mixture model is given by

$$\hat{p}(y_{n+m}|\mathbf{y}) = \frac{1}{M} \sum_{l=1}^M \sum_{j=1}^{r^{(l)}} \pi_{jrp}(t_{n+m}|\boldsymbol{\delta}_{rp}^{(l)})p(y_{n+m}|\mathbf{y}, \boldsymbol{\omega}_{jrp}^{(l)}), \quad (2)$$

where M is the number of iterates used, and t_{n+m} is the time corresponding to y_{n+m} . To quantify the distance between a known normal predictive density $p(y_{n+m}|\mathbf{y}, \boldsymbol{\omega})$ and its estimate $\hat{p}(y_{n+m}|\mathbf{y})$ based on (2), we use the Kullback-Leibler (KL) distance, given by

$$KL(\hat{p}(y_{n+m}|\mathbf{y}), p(y_{n+m}|\mathbf{y}, \boldsymbol{\omega})) = \int p(y_{n+m}|\mathbf{y}, \boldsymbol{\omega}) \log \frac{\hat{p}(y_{n+m}|\mathbf{y})}{p(y_{n+m}|\mathbf{y}, \boldsymbol{\omega})} dy_{n+m}. \quad (3)$$

This distance satisfies $KL(\hat{p}, p) \leq 0$ with equality if and only if the two densities are equal. For a normal density $p(y_{n+m}|\mathbf{y}, \boldsymbol{\omega})$, the integral in (3) can be approximated by

$$KL_{MC}(\hat{p}(y_{n+m}|\mathbf{y}), p(y_{n+m}|\mathbf{y}, \boldsymbol{\omega})) = \frac{1}{K} \sum_{i=1}^K g(y_{n+m}^{(i)}), \quad (4)$$

where $y_{n+m}^{(i)}$ is drawn from $p(y_{n+m}|\mathbf{y}, \boldsymbol{\omega})$, $i = 1, \dots, K$.

5 Simulation

We generated datasets of length 2000 from the following piecewise autoregressive model

$$y_t = \begin{cases} \sum_{k=1}^6 \phi_{k1}y_{t-k} + \sigma_1\epsilon_t^{(1)} & \text{for } 1 \leq t \leq 200 \\ \sum_{k=1}^6 \phi_{k2}y_{t-k} + \sigma_2\epsilon_t^{(2)} & \text{for } 201 \leq t \leq 1000 \\ \sum_{k=1}^6 \phi_{k3}y_{t-k} + \sigma_3\epsilon_t^{(3)} & \text{for } 1001 \leq t \leq 1300 \\ \sum_{k=1}^6 \phi_{k4}y_{t-k} + \sigma_4\epsilon_t^{(4)} & \text{for } 1301 \leq t \leq 1600 \\ \sum_{k=1}^6 \phi_{k5}y_{t-k} + \sigma_5\epsilon_t^{(5)} & \text{for } 1601 \leq t \leq 2000, \end{cases}$$

with parameter values

j	ϕ_{1j}	ϕ_{2j}	ϕ_{3j}	ϕ_{4j}	ϕ_{5j}	ϕ_{6j}	σ_j
1	0.8874	-0.8523	0.2484	-0.6520	0.3224	-0.3287	0.0429
2	0.6955	-0.5518	0.3117	-0.6293	0.1137	-0.1003	0.0169
3	1.3415	-1.3702	0.8900	-0.9627	0.5807	-0.4173	0.0686
4	0.9776	-0.8560	0.4272	-0.6103	0.2016	-0.1631	0.0326
5	0.7995	-0.6821	0.2463	-0.5712	0.1656	-0.2169	0.0188

For each dataset, m -step-ahead forecasts were computed for $m = 1, \dots, 5$, using our mixture model once when r was allowed to be estimated and a second time when r was fixed at 1. For each m -step ahead forecast, the KL divergence between the true predictive distribution and the ones estimated based on the two model fits were computed. Although we generated 50 datasets, the following results are based on a single dataset, but similar results were obtained from the other datasets.

m	$r = 1$	$r > 1$
1	-61.5524	-1.8610
2	-44.9292	-2.0885
3	-37.6319	-2.2498
4	-35.3171	-2.3096
5	-37.3917	-2.4301

It is seen that when r is allowed to be estimated, the (absolute value of the) KL divergence is much smaller than the one when r is fixed at 1.

6 Example

Referring to the SOI data described above, Figure 2 presents the posterior distribution of the number of model components. The probabilities are about 0.44 and 0.56 for one and two components, respectively. Figure 3 presents the estimated mixing weights, $\hat{\pi}_{tj2}$, $j = 1, 2$, against time, in a mixture of two components. It is seen that two components are needed to capture the behavior of the time series at the beginning of the period.

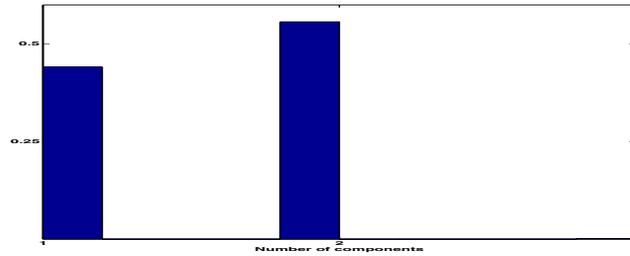
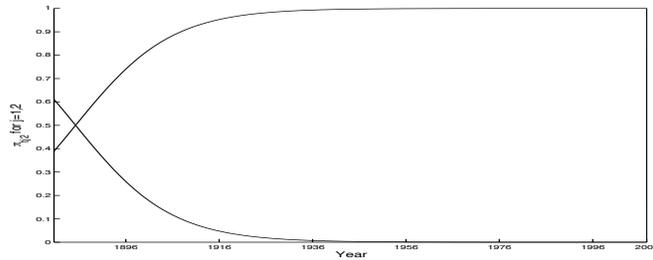


FIGURE 2. Posterior distribution of the number of components.

FIGURE 3. The estimated mixing weights, $\hat{\pi}_{tj2}$, $j = 1, 2$ in a mixture of two components.

References

- Davis, R.A., Lee, T.C.M. and Rodriguez-Yam, G.A. (2006). Structural breaks estimation for nonstationary time series models. *Journal of the American Statistical Association*, **101**, 223-239.
- Gerlach, R., Carter, C. and Kohn, R. (2000). Efficient Bayesian inference for dynamic mixture models. *Journal of the American Statistical Association*, **95**, 819-828.
- Green, P. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, **82**, 711-732.
- Jacobs, R.A., Jordan, M.I., Nowlan, S.J. and Hinton, G.E. (1991). Adaptive mixtures of local experts. *Neural Computation*, **3**, 79-87.
- Kitagawa, G. and Akaike, H. (1978). A procedure for the modeling of nonstationary time series. *Annals of the Institute of Statistical Mathematics*, **30**, 351-363.
- Lau, J.W. and So, M.K.P. (2008). Bayesian mixture of autoregressive models. *Computational Statistics and Data Analysis*, **53**, 38-60.

- Marin, J-M and C.P., Robert (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*, Springer.
- Ombao, H.C., Raz, J.A., Von Sachs, R. and Malow, B.A. (2001). Automatic statistical analysis of bivariate nonstationary time series. *Journal of the American Statistical Association*, **96**, 543-560.
- Prado, R. and Huerta, G. (2002). Time-varying autoregressions with model order uncertainty. *Journal of Time Series Analysis*, **23**, 599-618.
- Rosen, O., Stoffer, D.S. and Wood, S. (2009). Local spectral analysis via a Bayesian mixture of smoothing splines. *Journal of the American Statistical Association*, **104**, 249-262.
- Timmermann, A., Oberhuber, J., Bacher, A., Esch, M., Latif, M. and Roeckner, E. (1999). Increased El Niño frequency in a climate model forced by future greenhouse warming. *Nature*, **398**, 694-697.
- West, M., Prado, R. and Krystal, A. (1999). Evaluation and comparison of EEG traces: latent structure in non-stationary time series. *Journal of the American Statistical Association*, **94**, 1083-1095.

Assessment of variance components in elliptical nonlinear models for correlated data

Cibele M. Russo¹, Reiko Aoki² and Gilberto A. Paula¹

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 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 consider a score-type test for assessing variance components in nonlinear models with random effects assuming elliptical errors. The elliptical class includes light- and heavy-tailed distributions such as Student- t , power exponential, logistic, generalized Student- t , generalized logistic, contaminated normal, the normal itself, among others, and represents an alternative to the gaussian models in cases of heavy tails or extreme observations, for instance. Considering a score-type test proposed by Silvapulle and Silvapulle (1995), we compare the sensitivity of the score statistic under normal, Student- t and power exponential models for the kinetics data set discussed by Vonesh and Carter (1992).

Keywords: nonlinear models; elliptical distributions; hypothesis testing; variance components; score tests

1 Introduction

The vast majority of nonlinear models in the literature consider normal errors although this assumption may not be appropriate in several cases, as in the presence of heavy-tails or outliers. Some recent works can be found in the literature using elliptical distributions, however for nonlinear models few works can be found in this context. For linear models, recent references are the works by Osorio et al. (2007) and Savalli et al. (2006). Galea et al. (2005) developed local influence for symmetrical nonlinear models.

Let us consider \mathbf{y}_i an m_i -dimensional vector with $E(\mathbf{y}_i) = \boldsymbol{\mu}_i = \mathbf{f}(\mathbf{x}_i, \boldsymbol{\alpha})$, for $i = 1, \dots, n$. A possible mixed-effects model for \mathbf{y}_i , is given by

$$\mathbf{y}_i = \mathbf{f}(\mathbf{x}_i, \boldsymbol{\alpha}) + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \dots, n, \quad (1)$$

with $\mathbf{f}(\mathbf{x}_i, \boldsymbol{\alpha}) = (f(\mathbf{x}_{i1}, \boldsymbol{\alpha}), \dots, f(\mathbf{x}_{im_i}, \boldsymbol{\alpha}))^T$ being an m_i -dimensional nonlinear function of $\boldsymbol{\alpha}$, \mathbf{x}_i is a vector of explanatory variables values, \mathbf{Z}_i is a

matrix of known constants, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_p)^T$ a vector of unknown parameters and $\mathbf{b}_i = (b_{i1}, \dots, b_{ir})^T$ the random-effects coefficients. Vonesh and Carter (1992) considered the model defined in (1), with \mathbf{b}_i and $\boldsymbol{\epsilon}_i$ assumed to be independent and to follow a normal distribution. We generalize this model considering elliptical models, such that

$$\begin{bmatrix} \mathbf{y}_i \\ \mathbf{b}_i \end{bmatrix} \sim \text{El}_{m_i+r} \left\{ \begin{pmatrix} \mathbf{f}(\mathbf{x}_i, \boldsymbol{\alpha}) \\ \mathbf{0} \end{pmatrix}; \begin{bmatrix} \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \sigma^2 \mathbf{I}_{m_i} & \mathbf{Z}_i \mathbf{D} \\ \mathbf{D} \mathbf{Z}_i^T & \mathbf{D} \end{bmatrix} \right\}, \quad (2)$$

where the matrices $\boldsymbol{\Sigma}_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \sigma^2 \mathbf{I}_{m_i}$, \mathbf{D} , and $\mathbf{Z}_i \mathbf{D}$ are proportional to the variance-covariance matrices $\text{Var}(\mathbf{y}_i)$, $\text{Var}(\mathbf{b}_i)$ and $\text{Cov}(\mathbf{y}_i, \mathbf{b}_i)$, respectively, by a quantity $\zeta_i > 0$ which depends on the assumed elliptical distribution (under the normal case, $\zeta_i = 1$) (see Fang et al., 1990). We assume here that $\mathbf{D} = \mathbf{D}(\boldsymbol{\tau})$ is diagonal, which means that the random effects are uncorrelated. We work on the marginal model, namely $\mathbf{y}_i \sim \text{El}_{m_i}(\mathbf{f}(\mathbf{x}_i, \boldsymbol{\alpha}); \boldsymbol{\Sigma}_i)$, which preserves the mean of the hierarchical model without requiring numerical integration. The log-likelihood function $L_i = L_i(\mathbf{y}_i; \mathbf{x}_i, \boldsymbol{\alpha}, \boldsymbol{\Sigma}_i)$ is given by

$$L_i = -\frac{1}{2} \log |\boldsymbol{\Sigma}_i| + \log g \{ [\mathbf{y}_i - \mathbf{f}(\mathbf{x}_i, \boldsymbol{\alpha})]^T \boldsymbol{\Sigma}_i^{-1} [\mathbf{y}_i - \mathbf{f}(\mathbf{x}_i, \boldsymbol{\alpha})] \}, \quad (3)$$

with $g : \mathbb{R} \rightarrow [0, \infty)$ such that $\int_0^\infty u^{\frac{m}{2}-1} g(u) du < \infty$ is known as the density generator function.

2 Parameter estimation

The parameter vector to be estimated is given by $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \sigma^2, \boldsymbol{\tau}^T)^T$. Let $u_i = [(\mathbf{y}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)]$ be the Mahalanobis distance for each individual i , $i = 1, \dots, n$. Assuming g continuous and twice differentiable, it is usable in the elliptic context to define $W_g(u_i)$ and $W'_g(u_i)$ such that $W_g(u_i) = [d \log g(u_i)]/du_i$ and $W'_g(u_i) = [dW_g(u_i)]/du_i$. The total log-likelihood function for the marginal model is given by $L(\boldsymbol{\theta}) = \sum_{i=1}^n L_i(\mathbf{y}_i; \mathbf{x}_i, \boldsymbol{\theta})$ and the score functions may be written as

$$\begin{aligned} \mathbf{U}_\alpha &= \sum_{i=1}^n v_i \mathbf{J}_i^T \boldsymbol{\Sigma}_i^{-1} [\mathbf{y}_i - \mathbf{f}(\mathbf{x}_i, \boldsymbol{\alpha})], \text{ and} \\ \mathbf{U}_\sigma &= (U_{\sigma_1}, \dots, U_{\sigma_{r+1}})^T, \text{ with} \\ U_{\sigma_j} &= -\frac{1}{2} \sum_{i=1}^n \left\{ \text{tr} \left[\boldsymbol{\Sigma}_i^{-1} \dot{\boldsymbol{\Sigma}}_i(j) \right] - v_i \mathbf{r}_i^T \boldsymbol{\Sigma}_i^{-1} \dot{\boldsymbol{\Sigma}}_i^{-1}(j) \boldsymbol{\Sigma}_i^{-1} \mathbf{r}_i \right\}, \end{aligned}$$

in which $v_i = -2W_g(u_i)$, $\mathbf{r}_i = [\mathbf{y}_i - \mathbf{f}(\mathbf{x}_i, \boldsymbol{\alpha})]$, $\mathbf{J}_i = \partial \mathbf{f}(\mathbf{x}_i, \boldsymbol{\alpha}) / \partial \boldsymbol{\alpha}^T$, $\dot{\boldsymbol{\Sigma}}_i(j) = \partial \boldsymbol{\Sigma}_i / \partial \sigma_j$, for $j = 1, \dots, r + 1$, $i = 1, \dots, n$, and $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_{r+1})^T = (\sigma^2, \boldsymbol{\tau}^T)^T$. The quantity v_i that appears in the score functions for $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$ may be interpreted as a weight which, in particular, is inversely proportional to the Mahalanobis distance for Student-t and power exponential

distribution ($0 < \lambda < 1$). So, larger values for u_i lead to smaller values for v_i , which means a kind of qualitative robustness for the maximum likelihood estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\sigma}}$ as observed by Lucas (1997) for univariate Student-t models. The Fisher information matrix for $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\sigma}^T)^T$ is given by the block diagonal matrix $\mathbf{K}_{\boldsymbol{\theta}\boldsymbol{\theta}} = \text{diag}\{\mathbf{K}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}, \mathbf{K}_{\boldsymbol{\sigma}\boldsymbol{\sigma}}\}$, in which

$$\begin{aligned} \mathbf{K}_{\boldsymbol{\alpha}\boldsymbol{\alpha}} &= \sum_{i=1}^n \frac{4d_{g_i}}{m_i} \mathbf{J}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{J}_i \text{ and} \\ \mathbf{K}_{\boldsymbol{\sigma}\boldsymbol{\sigma}} &= \sum_{i=1}^n K_{i\sigma}, \text{ whose } qs\text{-th element is given by} \\ K_{i\sigma,qs} &= \frac{a_{qsi}}{4} \left(\frac{4f_{g_i}}{m_i(m_i + 2)} - 1 \right) + \frac{2f_{g_i}}{m_i(m_i + 2)} \text{tr} \left[\boldsymbol{\Sigma}_i^{-1} \dot{\boldsymbol{\Sigma}}_i(r) \boldsymbol{\Sigma}_i^{-1} \dot{\boldsymbol{\Sigma}}_i(s) \right], \end{aligned}$$

with $d_{g_i} = E \{W_g^2(U_i)U_i\}$, $f_{g_i} = E \{W_g^2(U_i)U_i^2\}$ where $U_i = \|\mathbf{Z}_i\|^2$, $\mathbf{Z}_i \sim \text{El}_{m_i}(0, \mathbf{I}_{m_i}, g)$ and $a_{qsi} = \text{tr} \left[\boldsymbol{\Sigma}_i^{-1} \dot{\boldsymbol{\Sigma}}_i(r) \right] \text{tr} \left[\boldsymbol{\Sigma}_i^{-1} \dot{\boldsymbol{\Sigma}}_i(s) \right]$.

An iterative algorithm to obtain the maximum likelihood estimates for $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$ using the Fisher scoring method is given by

$$\begin{aligned} \boldsymbol{\alpha}^{(m+1)} &= \boldsymbol{\alpha}^{(m)} + \mathbf{K}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^{(-m)} \mathbf{U}_{\boldsymbol{\alpha}}^{(m)} \text{ and} \\ \boldsymbol{\sigma}^{(m+1)} &= \boldsymbol{\sigma}^{(m)} + \mathbf{K}_{\boldsymbol{\sigma}\boldsymbol{\sigma}}^{(-m)} \mathbf{U}_{\boldsymbol{\sigma}}^{(m)}, \quad m = 0, 1, 2, \dots \end{aligned}$$

with $\mathbf{K}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}$, $\mathbf{U}_{\boldsymbol{\alpha}}$, $\mathbf{K}_{\boldsymbol{\sigma}\boldsymbol{\sigma}}$ and $\mathbf{U}_{\boldsymbol{\sigma}}$ as presented above. The initial values for the algorithm can be the least squares estimates, and the estimates of the random effects can be obtained by using the empirical Bayes method.

3 Assessing variance components

Considering the score-type test proposed by Silvapulle and Silvapulle (1995) and also discussed by Savalli et al. (2006), we have evaluated the hypothesis tests for the absence of variance components in model (1) by using the marginal model for which the regularity conditions presented by Silvapulle and Silvapulle (1995) seem reasonable to be satisfied. Assuming $\mathbf{D} = \text{diag}\{\tau_1, \dots, \tau_r\}^T$, we consider unilateral tests given by $H_0 : \boldsymbol{\tau} = \mathbf{0}$ against $H_1 : \boldsymbol{\tau} > \mathbf{0}$, with at least one strict inequality in H_1 . Consider the partitions in \mathbf{U} and $\mathbf{K}_{\boldsymbol{\theta}\boldsymbol{\theta}}$ according to $(\boldsymbol{\lambda}^T, \boldsymbol{\tau}^T)^T$, with $\boldsymbol{\lambda} = (\boldsymbol{\alpha}^T, \sigma^2)^T$, resulting in $\mathbf{U} = (\mathbf{U}_{\boldsymbol{\lambda}}^T, \mathbf{U}_{\boldsymbol{\tau}}^T)^T$ and $\mathbf{K}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}$, $\mathbf{K}_{\boldsymbol{\lambda}\boldsymbol{\tau}}$, $\mathbf{K}_{\boldsymbol{\tau}\boldsymbol{\lambda}}$ and $\mathbf{K}_{\boldsymbol{\tau}\boldsymbol{\tau}}$. Also, construct $\mathbf{Z} = [\mathbf{U}_{\boldsymbol{\tau}} - \mathbf{K}_{\boldsymbol{\tau}\boldsymbol{\sigma}} \mathbf{K}_{\boldsymbol{\sigma}\boldsymbol{\sigma}}^{-1} \mathbf{U}_{\boldsymbol{\sigma}}]$, so that \mathbf{Z} and $\mathbf{U}_{\boldsymbol{\lambda}}$ are independent. Thus the score-type statistic, T_S , is such that

$$T_S = \tilde{\mathbf{Z}}^T \tilde{\mathbf{K}}^{\boldsymbol{\tau}\boldsymbol{\tau}} \tilde{\mathbf{Z}} - \inf_{a \geq 0} \{(\tilde{\mathbf{Z}} - a)^T \tilde{\mathbf{K}}^{\boldsymbol{\tau}\boldsymbol{\tau}} (\tilde{\mathbf{Z}} - a)\},$$

where $\tilde{\mathbf{K}}^{\boldsymbol{\tau}\boldsymbol{\tau}} = \text{Var}(\hat{\boldsymbol{\tau}})$ is obtained partitioning $\mathbf{K}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}$ and all the parameters are evaluated under the null hypothesis $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\alpha}}^T, \tilde{\sigma}^2, \mathbf{0}^T)^T$. Under

regularity conditions, $T_S \stackrel{H_0}{\sim} \sum_{\ell=0}^r \omega(\ell, \mathbf{\Delta}) \chi_{\ell}^2$ follows a mixture of central chi-squared distributions in which χ_0^2 denotes the degenerated distribution at the origin, $\mathbf{\Delta} = \text{Var}(\hat{\boldsymbol{\tau}})$ and $\omega(\ell, \mathbf{\Delta})$'s are known as level probabilities and are expressed as functions of the coefficients of linear correlation associated with the $r \times r$ matrix $\mathbf{\Delta}$. If $\mathbf{\Delta}$ is a diagonal matrix, which signifies asymptotic independence between $\hat{\tau}_1, \hat{\tau}_2, \dots, \hat{\tau}_r$, then the level probabilities assume a recursive form $\omega(\ell, \mathbf{\Delta}) = \binom{r}{\ell} 2^{-r}, \ell = 0, \dots, r$ (more details in Savalli et al., 2006). Another test that is considered is given by $H_0 : \boldsymbol{\tau}_1 = \mathbf{0}$ against $H_1 : \boldsymbol{\tau}_1 > \mathbf{0}$ and/or $\boldsymbol{\tau}_2 > \mathbf{0}$, with $\boldsymbol{\tau} = (\boldsymbol{\tau}_1^T, \boldsymbol{\tau}_2^T)^T$, which means to test if the random effects related to $\boldsymbol{\tau}_1$ are not significant given $\boldsymbol{\tau}_2$ in the model.

4 Application

Vonesh & Carter (1992) presented a model for hemodialysis longitudinal data, relating the ultrafiltration rate (UFR, in ml/hr) at which water is removed and the transmembrane pressure (TMP, in mmHg) that is exerted on the dialyzer membrane in the water transport kinetics characterized by the nonlinear relationship

$$\text{UFR} = \alpha_0 \{1 - \exp[-\alpha_1(\text{TMP} - \alpha_2)]\},$$

where α_0 represents the maximum UFR one can attain due to protein polarization, α_1 is a hydraulic permeability transport rate, and α_2 is the transmembrane pressure required to offset patient oncotic pressure. In an experiment the ultrafiltration rate was measured in seven different transmembrane pressure levels in 20 high flux membrane dialyzers *in vitro* using bovine blood, with blood flow rates $Qb = 200\text{ml/min}$ ($\gamma_1 = 1$) and $Qb = 300\text{ml/min}$ ($\gamma_2 = 0$), which leads to consider $\alpha_0 = \alpha_{01}\gamma_i + \alpha_{02}(1 - \gamma_i)$, $\alpha_1 = \alpha_{11}\gamma_i + \alpha_{12}(1 - \gamma_i)$ and $\alpha_2 = \alpha_{21}\gamma_i + \alpha_{22}(1 - \gamma_i)$, $i = 1, 2$. Regarding the elliptical nonlinear model given by (1) and (2) using the marginal model, we have considered the model fitted by Vonesh & Carter (1992), with $\mathbf{Z}_i = [\partial f(\mathbf{x}_{ij}, \boldsymbol{\alpha})/\partial \alpha_0, \partial f(\mathbf{x}_{ij}, \boldsymbol{\alpha})/\partial \alpha_1]_{\boldsymbol{\alpha}=\bar{\boldsymbol{\alpha}}}$ in which $\bar{\boldsymbol{\alpha}}$ denotes the ordinary least squares estimate of $\boldsymbol{\alpha}$. The scale matrix \mathbf{D} , which is proportional to the variances of the two random effects \mathbf{b}_1 and \mathbf{b}_2 , is given by $\mathbf{D} = \text{diag}(\tau_1, \tau_2)$.

The standardized residuals under the normal model can be observed in Figure 1, in which the largest residuals in absolute value belong to the observations 10.6, 4.4, 10.7, 12.7, 7.4, 8.7, 1.5, 18.4 and 17.4 (dialyzer.measure). Considering the blood flow rate $Qb = 200\text{ml/min}$, dialyzers 10, 4, 7 and 8 presented the greatest decreases in the value of UFR from the fifth to the sixth, forth to the fifth, forth to the fifth and sixth to the seventh measures, respectively, dialyzer 10 also has the smallest sixth and seventh measures

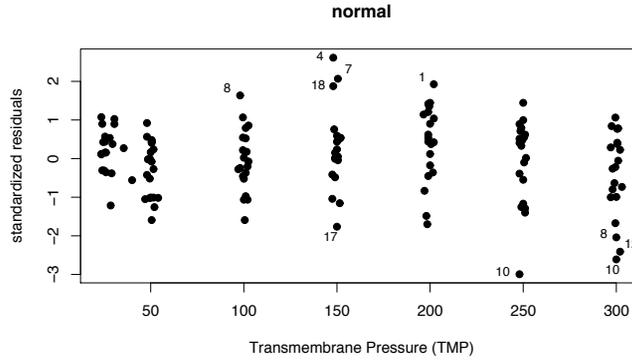


FIGURE 1. Standardized residuals under the normal model.

TABLE 1. Obtained values of T_S statistic and p-values considering the test of hypothesis $H_0 : \tau_2 = 0$ against $H_1 : \tau_2 > 0$ given τ_1 in the model.

	Normal		Student-t	
	T_S	p-value	T_S	p-value
All data	3.495	0.031	6.195	< 0.01
Removing dialyzer 13	2.474	0.058	4.716	0.015
Removing dialyzer 18	2.310	0.064	4.294	0.019
Removing dialyzer 20	2.251	0.067	3.059	0.040
Removing dialyzer 12 and 20	2.337	0.063	3.493	0.031
Removing dialyzer 12 and 18	1.759	0.092	3.865	0.025
Removing dialyzer 5 and 18	2.083	0.075	4.172	0.021

and dialyzer 1 has the largest fifth measure. Considering the blood flow rate $Qb = 300\text{ml/min}$, dialyzer 12 has the largest decrease from the sixth to the seventh measures while dialyzer 17 has the smallest fourth measure and finally dialyzer 18 has the largest fourth measure.

We have fitted Student-t and power exponential models to compare with the normal model fitted by Vonesh and Carter (1992). The degrees of freedom parameter for Student-t model ($\hat{\nu} = 8.7$) and the shape parameter for power exponential model ($\hat{\lambda} = 0.52$) were obtained using AIC and BIC information criteria. Also, the resulting values of AIC and BIC considering the three fitted models lead us to choose the Student-t model as the best fit.

Applying the results for the variance components score-type tests and considering the normal, Student-t and power exponential models, we have ob-

served that the score statistic is more sensitive in the normal model than in the Student-t and power exponential cases, producing inferential changes when some dialyzers were deleted. The null hypotheses $H_0 : \tau = 0$ (against $H_1 : \tau > 0$), $H_0 : \tau_1 = 0$ (against $H_1 : \tau_1 > 0$ given τ_2 in the model) and $H_0 : \tau_2 = 0$ (against $H_1 : \tau_2 > 0$ given τ_1 in the model) are rejected by the three considered models at the significance level of 5%, but the score test statistic for the last hypothesis showed to be sensitive to the exclusion of some dialyzers in the normal case, as it can be seen in Table 1. More specifically, the individual exclusion of the dialyzers 13, 18 and 20 induces changes at the significance level of 5% in the conclusion of the test under the normal model, as well as the joint exclusion of dialyzers 12 and 20; 12 and 18; 5 and 18. Similar results were obtained considering the power exponential model. These results are expected since the quantities v_i 's also appear as weights in the score statistic T_S .

5 Sensitivity study

To complement the results obtained by the hypotheses testing for the hemodialysis data, we analysed the sensitivity of the score statistic for the fitted models. That is, we perturbed the normal, Student-t and power exponential models and computed the relative changes in the score statistic due to the perturbation. A similar sensitivity study on the maximum likelihood estimates in symmetrical linear models has been performed by Cysneiros and Paula (2005).

We consider a response perturbation scheme, given by

$$\mathbf{y}_i = \mathbf{y}_i + \delta_i s_y, i = 1, \dots, n,$$

where s_y is the standard deviation of the vector \mathbf{y} and $\boldsymbol{\delta} = (\delta_1^T, \dots, \delta_n^T)^T$ is a perturbation vector, in which the no perturbation vector is clearly given by $\boldsymbol{\delta}_0 = \mathbf{0}$. Then, we define the quantity

$$W(\boldsymbol{\delta}) = \frac{|T_S(\boldsymbol{\delta}_0) - T_S(\boldsymbol{\delta})|}{T_S(\boldsymbol{\delta}_0)},$$

where $T_S(\boldsymbol{\delta})$ and $T_S(\boldsymbol{\delta}_0)$ represent the score statistic computed under the model perturbed by $\boldsymbol{\delta}$ and under the unperturbed model, respectively.

The chosen perturbation vector were based on observations with largest residuals. As the largest positive and negative residuals are found in dialyzer 4 and dialyzer 10, respectively, we considered $\boldsymbol{\delta}_4 = \mathbf{1}_{m_4} \delta$ and $\boldsymbol{\delta}_{10} = -\mathbf{1}_{m_{10}} \delta$, where $\mathbf{1}_{m_i}$ represents a $(m_i \times 1)$ vector of ones, $i = 4, 10$. The graphics of $W(\delta)$ against $\delta \in [0, 2]$, considering the score-type tests for $H_0 : \tau = 0$ against $H_1 : \tau > 0$ (Figure 2, left), $H_0 : \tau_1 = 0$ against $H_1 : \tau_1 > 0$ given τ_2 in the model (Figure 2, middle) and $H_0 : \tau_2 = 0$ against $H_1 : \tau_2 > 0$ given τ_1 in the model (Figure 2, right), show that the score statistic seems

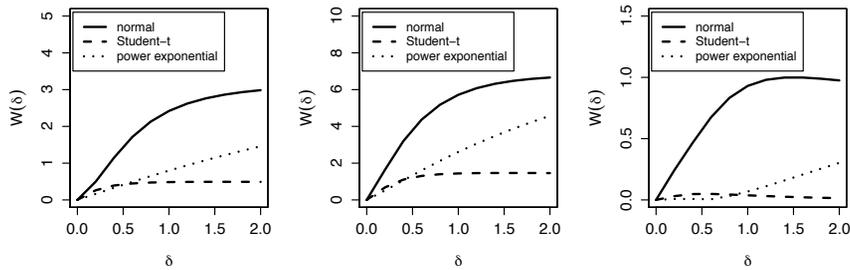


FIGURE 2. Behaviour of the score statistic under response perturbation.

to be more sensitive to the perturbations under the normal model than under the Student-t and power exponential models. Moreover, the score statistic under Student-t model appears to be the less sensitive to response perturbation among the fitted models.

Acknowledgments: The authors are grateful to CNPq and FAPESP, Brazil, which supported this research.

References

- Cysneiros, F. J. A. and Paula, G.A. (2005). Restricted methods in symmetrical linear regression models. *Computational Statistics and Data Analysis*, **49**, 689–708.
- Fang, K. T., Kotz, S. and Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions*. Chapman & Hall.
- Galea, M., Paula, G. A. and Cysneiros, F. J. (2005). On diagnostics in symmetrical nonlinear models. *Statistics & Probability Letters*, **73**, 459–467.
- Lucas, A. (1997). Robustness of the student t based M-estimator. *Communications in Statistics, Theory and Methods*, **26**, 1165–1182.
- Osorio, F., Paula, G. A. and Galea, M. (2007). Assessment of local influence in elliptical linear models with longitudinal structure. *Computational Statistics and Data Analysis*, **51**, 4354–4368.
- Savalli, C., Paula, G. A. and Cysneiros, F. J. A. (2006). Assessment of variance components in elliptical linear mixed models, *Statistical Modelling*, **6**, 59–76.

Silvapulle, M. J. and Silvapulle, P. (1995). A score test against one-sided alternatives, *Journal of the American Statistical Association*, **90**, 342-349.

Vonesh, E. F. and Carter, R. L. (1992). Mixed-effects nonlinear regression for unbalanced repeated measures, *Biometrics*, **48**, 1-17.

Non-crossing smooth expectile curves

Sabine K. Schnabel^{1,2} and Paul H.C. Eilers³

¹ Max Planck Institute for Demographic Research, Rostock, Germany

² Biometris, Wageningen University, P.O. Box 100, 6700 AC Wageningen, The Netherlands; schnabel@demogr.mpg.de (communicating author)

³ Erasmus Medical Center, Department of Biostatistics, Postbus 2040, 3000 CA Rotterdam, The Netherlands; p.eilers@erasmusmc.nl

Abstract: As an alternative to quantile smoothing we propose expectile smoothing using least asymmetrically weighted squares. The problem of crossing curves will be addressed by an explicit bundle using a bilinear form which allows also for direct estimation of the underlying density.

Keywords: Expectiles, smoothing, crossing curves, density

1 Introduction

Quantile smoothing (QS) has become a popular tool (see Koenker, 2005) for studying both trend and spread in scatterplots. We try to achieve the same for expectile smoothing (ES) based on least asymmetrically weighted squares (LAWS) as introduced by Newey and Powell in 1987. Both QS and ES can result in crossing curves, especially for small data sets. The reason is that each curve is estimated independent of the others. Several authors have proposed solutions for this problem in QS. Among the suggested methods we find (natural) monotoneization of the empirical curves (Chernozhukov et al., 2007), non-parametric estimation (Dette and Volgushev, 2008), restricted regression quantiles (He, 1997), double kernel estimators (Yu and Jones, 1998) as well as a few other models.

We propose an explicit bundle model for expectile curves, consisting of three components: 1) a smooth trend, 2) a vector of standardized expectiles, which is modulated by 3) a smooth amplitude curve. The smooth curves are modeled by P -splines. The standardized expectiles represent a common conditional distribution that is shifted and scaled by the smooth trend and the amplitude. We also present an algorithm to compute a smooth (non-negative) density from a series of expectiles.

2 Least asymmetrically weighted squares

In this section we first describe the theoretical derivations behind least asymmetrically weighted squares estimation. Second we present the LAWS

bundle model to overcome the problem of crossing expectiles.

2.1 Introduction to LAWS

In general LAWS minimizes the following goal, with $0 < p < 1$:

$$S = \sum_i w_i (y_i - \mu(x_i, p))^2$$

with weights

$$w_i = \begin{cases} p & \text{if } y_i > \mu(x_i, p) \\ 1 - p & \text{if } y_i \leq \mu(x_i, p) \end{cases}, \quad (1)$$

where y_i is the response variable and $\mu(x_i, p)$ is the estimated value according to a statistical model. The obtained functions are called p -expectiles (Newey and Powell, 1987). It is easy to fit any LAWS model: iterate between weighted regression and re-compute the weights. The goal function is convex, therefore a unique minimum is guaranteed.

We combine LAWS and P -splines (Eilers and Marx, 1996) and define

$$\mu(x_i, p) = \sum_j b_{ij} a_j,$$

where $B = [b_{ij}]$ contains a generous number of B -splines, and a their coefficients. Each p -expectile is estimated separately. A difference penalty on a , with parameter λ , allows for continuous tuning of smoothness. To optimize λ either asymmetric cross-validation (ACV) or an adaptation of Schall's (1991) method can be used (see Schnabel and Eilers (2009) for details). An comparison for these smoothing criteria using the LIDAR data can also be found in (Schnabel and Eilers, 2008).

2.2 The LAWS bundle model

Expectile curves for a series of values of p often cross in small data sets. In theory this is not possible, but commonly encountered in practise due to sampling variation. An example is shown in Figure 1a for a subset of the LIDAR data set from the R package `semipar`.

We propose the LAWS bundle model to prevent intersecting expectiles. In this bilinear model the expectiles $\mu(x_i, p)$ are defined by

$$\mu(x_i, p) = b(x_i) + g(p)a(x_i),$$

where $b(x)$ is a common smooth trend, $g(p)$ represents standardized expectiles, and $a(x)$ represents the local width of the bundle. The smooth functions of x are modeled by P -splines. To make the model identifiable, the constraint on the coefficients in the smooth function of the local width of the model $\sum_i^m (c_i^a)^2 = m$ is introduced.

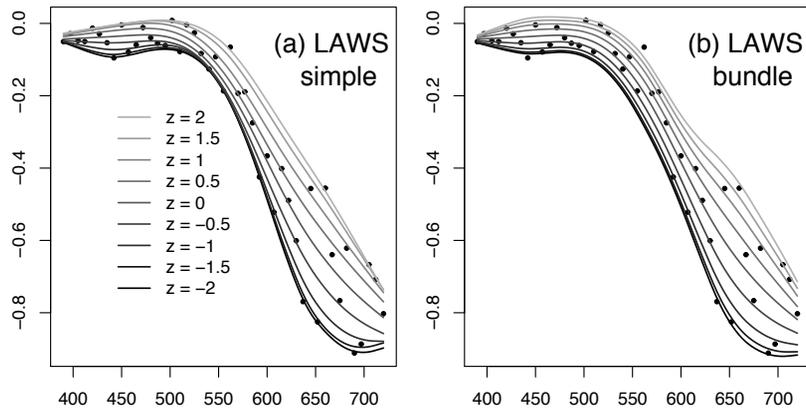


FIGURE 1. Estimated curves from a subset of the LIDAR data (45 observations). For better comparability expectiles are plotted for a scale of z expressed in standard deviations of $N(0, 1)$ which translates into p effectively ranging from 0.004 to 0.996. (a) Estimated expectiles using LAWS with the smoothing parameter determined by Schall's algorithm. (b) Estimated expectiles using the LAWS bundle model and 10-fold cross-validation.

The estimation procedure consists of two steps: in the first step the common trend $b(x)$ is estimated as a P -spline. In the second step the standardized expectiles $g(p)$ and the function describing the width of the bundle $a(x)$ are determined iteratively from the residuals estimated in step 1. To this end the following objective function is minimized:

$$S_B = \sum_{i=1}^m \sum_{j=1}^n w_{ij} [y_i - (b(x_i) + g_j a(x_i))]^2,$$

where the vector g with elements $g_j = g(p_j)$, $j = 1, \dots, J$ contains the standardized expectiles and w_{ij} is the asymmetric weight that follows from the sign of $y_i - (b(x_i) + g_j a(x_i))$ and p_j as defined in (1). We use a series of values for p_j , e.g. the same values that are used for the computation of individual expectile curves.

This set-up involves the choice of two smoothing parameters (λ_b, λ_a) . We use 10-fold asymmetric cross-validation and a grid search. Figure 1b shows the results.

We can plot g against p , similar to the quantiles. Alternatively we can plot it against theoretical expectiles of one or more reference distributions (like normal or uniform). This will give us an E-E plot analogous to the Q-Q plot. An example for a normal E-E plot is shown in Figure 2b.

2.3 Density estimation

It is possible to compute a density estimate from the vector of the standardized expectiles g . Suppose we approximate the density by a discrete vector f on a grid u . Then we have, for each p_j :

$$\sum_k w_{kj}(u_k - g_j)f_k = 0,$$

where $w_{kj} = p_j$ if $u_k > g_j$ and $w_{kj} = 1 - p_j$ otherwise. We want $\sum_k w_{kj}f_k(u_k - g_j)^2$ to be minimal. If we set $v_{kj}^* = w_{kj}(u_k - g_j)$, we have that $V^*f = 0$. In addition f has to be a proper discrete distribution, hence $\sum_k f_k = 1$. All conditions can be combined by adding a row of ones to V^* , giving V , and forming the vector r , consisting of J zeros and 1 one. The conditions can then be written as $Vf = r$.

To estimate f we form the objective function

$$S_D = \|Vf - r\|_2^2 + \lambda_f \|Df\|_2^2,$$

where $\|\cdot\|_2$ is the $L2$ norm and D is a matrix that forms third differences. The first term of S_D contains the conditions, which should be respected. The second term is a roughness penalty. It has two goals: 1) to remove a possible ill condition, because f has more elements than the number of conditions and 2) to make f smooth.

There is no guarantee that all elements of f will be non-negative, and indeed we get negative estimates for the LIDAR data. As a remedy we set $f_k = e^{\eta_k}$, with the roughness penalty on η_k . The modified goal can be optimized by Newton-Raphson iterations. Details will be reported elsewhere.

Figure 2a shows the estimated densities for the LIDAR data using the two algorithms.

The weight of the penalty (λ_f) has been chosen subjectively. Optimizing the penalty is a subject for further research. In standard density smoothing we have cross-validation or AIC as guiding principles, but it is not yet clear whether they apply here.

An interesting next step might be to compute (standardized) quantiles by integrating the estimated density. We will not pursue that here.

2.4 Robustness

It seems reasonable to assume that expectiles are more efficient than quantiles. The latter only use the signs of residual, while the former use signs and size. Especially for small data sets this might be important. We have not yet tried to quantify efficiency gains. An interesting subject for future research is to compare, in simulations, quantiles as derived from the density estimate described above with those obtained from quantile regression.

The price of larger efficiency is reduced robustness against outliers. An example for this issue is shown in (Schnabel and Eilers, 2009b) where we also proposed an approach for detecting extreme observations.

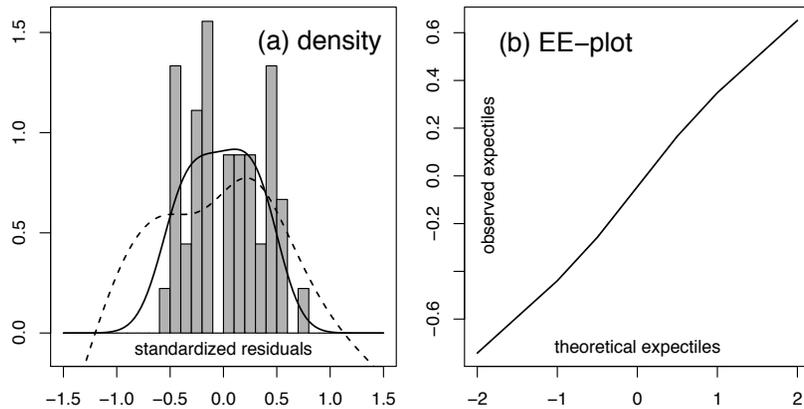


FIGURE 2. (a) Estimated densities from expectiles. Dashed line: basic algorithm; solid line: with constraints $f_k \geq 0$. (b) Expectile-expectile plot of the observed expectiles versus the theoretical expectiles of the standard normal distribution.

3 Applications

We used a every fifth observation in the well-known LIDAR data set from the `semipar` package for `R` as an example for the application of the proposed methods. This data is described in (Ruppert et al., 2003).

To be able to compare expectiles with e.g. quantiles, we need a common scale. For that we use a the standard normal distribution with cumulative distribution $F(z)$. We chose a “nice” grid for z (e.g. $z \in (-2\sigma, 2\sigma)$), and match $q = F(z)$ (for the quantile) with the value of the asymmetry parameter p that gives z as expectile.

The estimated curves using the bundle model are shown in Figure 1b and we see that the curves are smooth and do not intersect.

From the results of the bundle model we can estimate the underlying density as described above (see Figure 2a). On the abscissa are the standardized residuals $(y - b(x))/a(x)$. In order to get a better perspective the density can also be depicted together with the data as illustrated in Figure 3. As expected from the nature of the data the density is narrow at lower values of x and spread more toward the higher end of the data range.

4 Discussion

The LAWS bundle model provides expectile curves by using a bilinear model in which a common distribution –described by expectiles– is shifted

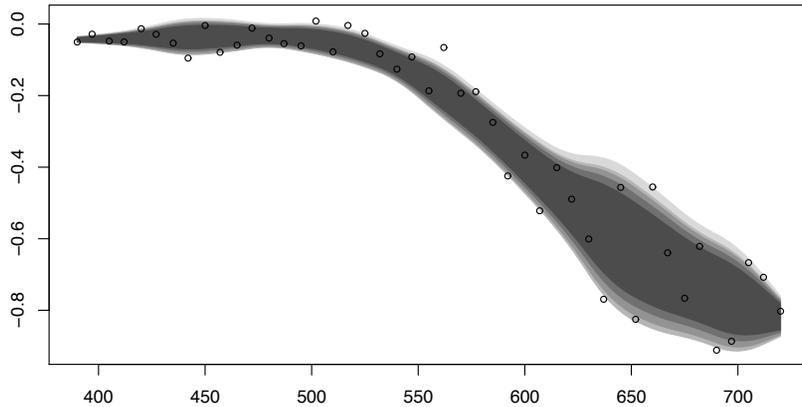


FIGURE 3. Subset of the LIDAR data and estimated underlying density. Darker areas indicate higher density, lighter areas indicate lower density.

and scaled. From the expectiles the distribution can be computed by a novel algorithm.

We note that the density estimation algorithm also works for standard expectiles (for any chosen value of x), as long as the curves do not cross. It might even be the case that crossing is not very harmful, because the objective function strikes a balance between respecting the observed expectiles and smoothness of the density. This will be part of our future research.

The common distribution is at the same time an advantage and a limitation. If the data fit in this harness, better results can be obtained because the degrees of freedom of the set of expectile curves is strongly reduced. But some data sets will be particular and then bias will occur. A challenge for further research is to develop ways to measure goodness of fit. We expect the bundle model to be particularly adequate for small data sets. We also note that the bundle model can be used to estimate smooth non-crossing quantile curves determined through the results of the model.

The proposed model can be used in a variety of different applications and settings. In our research on the relationship between life expectancy and economic production using data from a diverse set of countries, we found that expectiles are very useful to define frontiers and evaluate relative performance of individual countries (Schnabel and Eilers, 2009b). However, we also encountered problems with crossing expectile curves which can be easily solved by using the bundle model.

In our current research work we have extended the LAWS model by increasing the number of dimensions. Using tensor products of B -splines we aim at estimating so called smooth *expectile sheets* in the space spanned

by the independent variable and asymmetry parameter p . First analyses show very promising results. This approach adds another perspective to estimating expectiles.

References

- Chernozhukov, V., Fernandez-Val, I., and Galichon, A. (2007). Quantile and probability curves without crossing. MIT Department of Economics Working Paper 07-15, MIT.
- Dette, H., and Volgushev S. (2008). Non-crossing non-parametric estimates of quantile curves. *Journal of the Royal Statistical Society: Series B*, **70**, 609-627.
- Eilers, P.H.C., and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Sciences*, **11**, 89-121.
- He, X. (1997). Quantile curves without crossing. *The American Statistician*, **51**, 186-192.
- Koenker, R. (2005). *Quantile regression*. New York: Cambridge University Press.
- Newey, W.K., and Powell, J.L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, **55**, 819-847.
- Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. New York: Cambridge University Press.
- Schall, R. (1991). Estimation in Generalized Linear Models with random effects. *Biometrika*, **78**, 719-727.
- Schnabel, S.K., and Eilers, P.H.C. (2008). Optimal Expectile Smoothing. In: Proceedings of the 23rd International Workshop on Statistical Modelling. Utrecht, The Netherlands.
- Schnabel, S.K., and Eilers, P.H.C. (2009a). Optimal Expectile Smoothing. *To appear*.
- Schnabel, S.K., and Eilers, P.H.C. (2009b). An analysis of life expectancy and economic production using expectile frontier zones. *To appear*.
- Yu, K., and Jones, M. C. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, **93**, 228-237.

A Regression Model for Count Data with Observation-Level Dispersion

Kimberly F. Sellers¹ and Galit Shmueli²

¹ 306 St. Mary's Hall; Department of Mathematics; Georgetown University; Washington, DC 20057; kfs7@georgetown.edu

² Dept of Decision, Operations & Information Technologies; 4361 Van Munching Hall; Smith School of Business; University of Maryland; College Park, MD 20742; gshmueli@rhsmith.umd.edu

Abstract: While Poisson regression is a popular tool for modeling count data, it is limited by its associated model assumptions. One assumption is that the response variable follows a Poisson distribution. However, over- or under-dispersion are common in practice and are not accommodated by Poisson regression. In addition, the dispersion is assumed fixed across observations, whereas in practice dispersion may vary across groups or according to some other factor. Recently, Sellers and Shmueli (2008) introduced the Conway-Maxwell-Poisson (CMP) regression, based on the CMP distribution. CMP regression generalizes both Poisson and logistic regression models and allows for over- or under-dispersed count data. The model structure introduced, however, assumes a fixed dispersion level across all observations. In this paper, we extend the CMP regression model to account for observation-level dispersion. We discuss model estimation, inference, diagnostics, and interpretation, and present a variable selection technique. We then compare our model to several alternatives and illustrate its advantages and usefulness using datasets with varying types and levels of dispersion.

Keywords: Conway-Maxwell Poisson distribution; generalized linear models (GLM); generalized Poisson; observation-level (varying) dispersion.

1 Introduction

Poisson regression models are most widely used to model relationships in count data; however, the model assumption [$\text{Var}(Y_i) = \text{E}(Y_i)$] is limiting. More generally, data exhibit over- or under-dispersion. Several papers offer ways to circumvent this problem, most of which focus on addressing the matter of overdispersion (McCullagh and Nelder, 1997; Famoye, 1993). Recently, Sellers and Shmueli (2008) introduced the CMP regression model, based on the Conway-Maxwell-Poisson (CMP) distribution, which allows handling over- and under-dispersed data. CMP regression also generalizes Poisson regression and logistic regression. Although it offers flexibility in terms of dispersion, it assumes that the associated level of dispersion is constant across all observations. We term this model "constant-dispersion

CMP regression". In this paper, we propose an extension of the CMP regression model which allows for observation-level dispersion. We start by describing the CMP distribution in Section 1.1, and then the constant-dispersion regression model by Sellers and Shmueli (2008) in Section 1.2. Approaching the CMP distribution from a GLM perspective, we use $\log \lambda$ as the link function and show its benefits with regard to estimation and inference. Section 2 introduces "observation-level-dispersion CMP regression", generalizing the ideas expressed in the previous section to consider a variable dispersion parameter and thus modeling the dispersion as a function of the explanatory variables. In Section 3 we describe a hypothesis testing procedure to determine the appropriateness of using a CMP regression model that allows for constant- or observation-level dispersion, and consider a variable selection construct that accounts for all subsets of main effects associated with the data relationship and the associated dispersion. Section 4 illustrates the application of the observation-level dispersion CMP regression. We compare the results of the constant- versus observation-level dispersion CMP regression in terms of fit, inference, and interpretation, and discuss the variable selection results as they relate to the examples provided.

1.1 The CMP Distribution

The CMP probability distribution function takes the form

$$P(Y_i = y_i) = \frac{\lambda^{y_i}}{(y_i!)^\nu Z(\lambda_i, \nu)}, \quad y_i = 0, 1, 2, \dots, \quad i = 1, \dots, n, \quad (1)$$

for a random variable Y_i , where $Z(\lambda_i, \nu) = \sum_{s=0}^{\infty} \frac{\lambda_i^s}{(s!)^\nu}$. In this setting, $\lambda_i = E(Y_i^\nu)$, while ν is the dispersion parameter. The CMP distribution includes three well-known distributions as special cases: Poisson ($\nu = 1$), geometric ($\nu = 0, \lambda_i < 1$), and Bernoulli ($\nu \rightarrow \infty$ with probability $\frac{\lambda_i}{1+\lambda_i}$). See Shmueli et al. (2005) for details regarding this distribution.

1.2 CMP Model estimation with constant dispersion

Sellers and Shmueli (2008) took a GLM approach and used the link function $\eta(E(Y)) = \log \lambda$ that indirectly models the relationship between $E(\mathbf{Y})$ and $\mathbf{X}\beta = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$, that allows for estimating β and constant ν via the associated normal equations. Using the Poisson estimates, $\beta^{(0)}$ and $\nu^{(0)} = 1$ (or $\gamma^{(0)} = \ln \nu^{(0)} = 0$) as the starting values, these equations can be solved iteratively via an appropriate iterative reweighted least squares procedure to determine the maximum likelihood estimates for β and ν (or β and γ , respectively). The associated standard errors of the estimated coefficients are derived using the Fisher Information matrix; see Sellers

and Shmueli (2008) for details. *R* code for estimating the CMP regression coefficients and standard errors under the constant dispersion assumption is available at www9.georgetown.edu/faculty/kfs7/research.

2 CMP Model estimation with observation-level dispersion

We now allow for the dispersion parameter, ν_i , to vary with observation i , and consider a relationship between ν_i and the observations encapsulated in the $(p + 1)$ -dimensional row vector, \mathbf{X}_i . Accordingly, we write the log-likelihood for observation i as

$$\log L_i(\lambda_i, \nu_i | y_i) = y_i \log \lambda_i - \nu_i \log y_i! - \log Z(\lambda_i, \nu_i), \quad (2)$$

where

$$\log \lambda_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \doteq \mathbf{X}_i \beta, \text{ and} \quad (3)$$

$$\log \nu_i = \gamma_0 + \gamma_1 x_{i1} + \cdots + \gamma_p x_{ip} \doteq \mathbf{X}_i \gamma. \quad (4)$$

Since the CMP distribution belongs to the exponential family, we can determine appropriate normal equations for β and γ . Using the Poisson estimates, $\beta^{(0)}$ and $\gamma^{(0)} = 0$, as starting values, coefficient estimation can again be achieved via an appropriate iterative reweighted least squares procedure, or by using existing nonlinear optimization tools (e.g., `nlm` in *R*) to directly maximize the likelihood function. The associated standard errors of the estimated coefficients are derived in an analogous manner to that described in Sellers and Shmueli (2008).

3 Testing for Variable Dispersion, and Performing Variable Selection

Sellers and Shmueli (2008) established a hypothesis testing procedure to determine the need for using a CMP regression model over a simple Poisson regression model. In other words, they test whether $\nu = 1$ or not. We now ask the follow-up question: is the dispersion level fixed across observations, or is it dependent on one or more of the p covariates? More formally, we consider the set of hypotheses:

$$\begin{aligned} H_0 &: \gamma_i = 0 \text{ for } i = 1, \dots, p \quad \text{vs.} \\ H_1 &: \gamma_i \neq 0 \text{ for at least one } i \in \{1, \dots, p\}. \end{aligned} \quad (5)$$

The likelihood ratio statistic Λ and the derived test statistic C are given by

$$\Lambda = \frac{L(\hat{\beta}_{(0)}, \hat{\gamma}_{0(0)})}{L(\hat{\beta}, \hat{\gamma})} \quad (6)$$

$$C = -2 \log \Lambda = -2 \left[\log L \left(\hat{\beta}_{(0)}, \hat{\gamma}_{0(0)} \right) - \log L \left(\hat{\beta}, \hat{\gamma} \right) \right], \quad (7)$$

where $\hat{\gamma}_{i(0)} = 0$ for $i = 1, \dots, p$. $\hat{\beta}_{(0)}, \hat{\gamma}_{(0)}$ where $\nu_{(0)} = \exp(\mathbf{X}\gamma_{(0)})$ are the maximum likelihood estimates obtained under H_0 , i.e. they are the CMP estimates under the constant dispersion model; and $(\hat{\beta}, \hat{\gamma})$ are the maximum likelihood estimates under the variable dispersion model, obtained by Equations (3) and (4). Under the null hypothesis, C has an approximate χ^2 distribution with p degree of freedom. Therefore, we reject H_0 in favor of H_1 when $C > \chi_\alpha(p)$.

We have also created a variable selection procedure that considers all possible subsets and the associated Akaike Information Criterion corrected for small sample sizes, when necessary (AICc). Thus, we can use the AIC or AICc to determine the "best" subset for predicting the response from a set of predictors, whether using the constant or observation-level dispersion model framework.

4 Examples

We compare the constant and observation-level CMP regression models to datasets characterized by under- and over-dispersion. These datasets were analyzed in Sellers and Shmueli (2008), comparing the constant-dispersion CMP regression results to those from other potential regression models for the data, including Poisson, negative binomial (NB), linear with $\log(Y)$, restricted generalized Poisson (Famoye, 1993). In addition, we illustrate the variable selection procedure by applying it to the example datasets. We show how the procedure can be used to find the optimal subset of predictors for predictive purposes, as well as shedding light on the effect of different predictors on the dispersion.

4.1 An Under-dispersed Dataset

We consider the airfreight breakage example from Kutner et al. (2003), where data are given on 10 air shipments, each carrying 1000 ampules on the flight. For each shipment i , we have the number of times the carton was transferred from one aircraft to another (X_i) and the number of ampules found broken upon arrival (Y_i). A graphical representation of the data is provided in Figure 1.

Figure 1 illustrates the (potentially observation-level) dispersion present in the count data. Thus, we consider regression models that allow for a constant dispersion or an observation-level dispersion structure. Table 1 contains the parameter estimates determined under the respective models. The standard errors associated with the respective estimates, however, bring question to the statistically significant difference between the two models. We see that the estimates for γ_0 are somewhat similar between

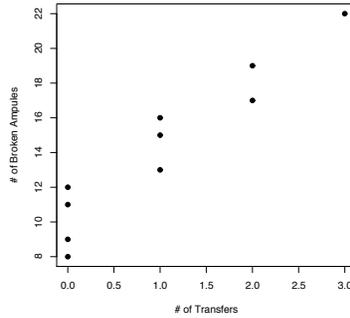


FIGURE 1. Scatter plot associated with airfreight breakage data.

TABLE 1. Estimated coefficients and standard errors (in parentheses) for the CMP regression models assuming fixed and variable dispersion for Airfreight example

Dispersion	$\hat{\beta}_0 (\hat{\sigma}_{\hat{\beta}_0})$	$\hat{\beta}_1 (\hat{\sigma}_{\hat{\beta}_1})$	$\hat{\gamma}_0 (\hat{\sigma}_{\hat{\gamma}_0})$	$\hat{\gamma}_1 (\hat{\sigma}_{\hat{\gamma}_1})$
Constant	13.8247 (6.2369)	1.4838 (0.6888)	1.7547 (0.954)	
Obs-level	15.5851 (0.7190)	4.6267 (0.3617)	1.8928 (0.3228)	0.1205 (0.1614)

the two models, and the estimate for γ_1 (in the observation-level dispersion model) has an associated standard error that allows for inclusion of 0 in the resulting confidence interval.

The dispersion test as described in Sellers and Shmueli (2008) yields a test statistic of $C = 9.10$ and associated p-value of 0.003, thus illustrating strong data (under-)dispersion and, therefore, a need to model the dataset with a more accommodating regression model, namely a CMP regression. Meanwhile, the hypothesis test for constant versus variable dispersion yields a test statistic of $C = 2.59$ and associated p-value of 0.11. Thus, one can argue that the dispersion does not statistically significantly vary by the number of aircraft transfers. We must, however, note the marginal p-value, and take the small sample size associated with this example into account. Thus, we consider variable selection under the observation-level dispersion structure to consider a broader realm of models.

Table 2 contains the variable selection results considering all relevant subsets for β and γ . Whether using the AIC or AIC_c result, we see that the best predictive models explain the number of broken ampules via the number of flight transfers. Recalling the marginal statistical insignificance noted in the above hypothesis test regarding the constant versus observation-level

TABLE 2. Variable selection results for airfreight example. We consider model selection with an observation-level dispersion assumption. The AIC and AIC_c values are provided for all models under consideration.

β_0	β_1	γ_0	γ_1	AIC	AIC_c
0.000	-0.019	0.000	-3.938	179.567	181.281
2.115	0.000	-0.223	0.000	60.897	62.611
13.825	1.484	1.755	0.000	43.290	47.290
13.294	0.000	1.711	-0.093	44.637	48.637
15.585	4.627	1.893	0.120	42.695	50.695

dispersion, this dataset's small sample size further accents this result. The observation-level dispersion model is found to be optimal via AIC, while the constant dispersion model is considered the best according to the corrected AIC result (AIC_c).

Given the small sample size here, the corrected AIC is more appropriate in this setting. Focusing our attention accordingly we see that, while the constant dispersion model produces the smallest AIC_c , the observation-level dispersion models produce associated AIC_c values that are close relative to that produced by the constant dispersion model that describes the number of broken ampules in relation to the number of freight transfers. Analogously, in consideration of the AIC results, we again see these three models all producing relatively similar AIC values. Thus, while the results in Table 2 imply that the constant dispersion model is optimal, one can argue for more data to better analyze this relationship.

4.2 An Over-dispersed Dataset

Lord et al. (2008) model crash data in 1995 at 868 signalized intersections located in Toronto, Ontario using a Bayesian formulation of a CMP regression for modeling the relationship between traffic variables and motor vehicle crashes. Sellers and Shmueli (2008) compare their CMP regression results to those of Lord et al. (2008) to find that the parameter estimates are identical under the constant dispersion construct, and compare them to other potential regression models. Meanwhile, Table 3 contains the resulting parameter estimates under the constant and variable dispersion models for comparison. We see here that the corresponding estimates for β and γ do not appear to be statistically significantly different, given their associated standard errors. We will pursue this hypothesis further in the hypothesis test for the existence of statistically significant variable dispersion.

The constant dispersion test yielded a test statistic of 518.37 with an associated p-value < 0.001 , thus noting the significant dispersion that exists in the data and thus the need to perform a CMP regression as opposed to a classical Poisson approach. Meanwhile, the test for variable dispersion yielded a test statistic of 1.02 with an associated p-value equaling 0.60.

TABLE 3. Estimated coefficients and standard errors (in parentheses) for the CMP regression models assuming fixed and variable dispersion

Dispersion	$\hat{\beta}_0 (\hat{\sigma}_{\hat{\beta}_0})$	$\hat{\beta}_1 (\hat{\sigma}_{\hat{\beta}_1})$	$\hat{\beta}_2 (\hat{\sigma}_{\hat{\beta}_2})$
Constant	-4.0862726 (0.2619)	0.2290205 (0.0216)	0.2762342 (0.0161)
Obs-level	-3.8390 (0.3858)	0.2340 (0.0661)	0.2444 (0.0324)
Dispersion	$\hat{\gamma}_0 (\hat{\sigma}_{\hat{\gamma}_0})$	$\hat{\gamma}_1 (\hat{\sigma}_{\hat{\gamma}_1})$	$\hat{\gamma}_2 (\hat{\sigma}_{\hat{\gamma}_2})$
Constant	-1.0522 (-3.8714)		
Obs-level	-0.7849 (0.2447)	0.0096 (0.0271)	-0.0385 (0.0064)

TABLE 4. Variable selection results for Toronto crash example. Note that we consider model selection with a constant dispersion assumption because the associated hypothesis test determined that a model with constant dispersion is adequate.

β_0	β_1	β_2	ν	AIC
0.055	0.000	0.000	0.049	6008.582
-1.806	0.000	0.262	0.271	5216.373
-1.527	0.166	0.000	0.094	5797.992
-4.086	0.229	0.276	0.349	5072.950

Given the large sample size, this demonstrates that the dispersion does not statistically significantly vary over predictor levels, and thus the assumption of a constant dispersion parameter is reasonable.

Because the constant dispersion model is sufficient, we pursue the question of variable selection under the constraint of constant dispersion. Table 4 provides the resulting parameter estimates and associated AIC for all possible subsets for consideration. As a result, the full model appears to provide the optimal choice for predicting the number of motor vehicle crashes, as demonstrated by the smallest AIC. All of the models have $\nu < 1$, which reflects the data overdispersion.

4.3 Summary

We have illustrated how the observation-level dispersion CMP model can be fitted to datasets with varying levels of dispersion. In both cases, statistical testing indicated that a constant dispersion level across all observations is better than a model that varies the dispersion level based on the covariates. We then used model selection to detect the predictor combination that yields the best predictive model. For the first example, this proved to be quite interesting because the marginal p-value associated with the hypothesis test followed with variable selection options that provided close AIC_c results where constant- or observation-level dispersion models could seem reasonable.

References

- Famoye, F. (1993) Restricted generalized Poisson regression model. *Communications in Statistics - Theory and Methods*, **22**(5), 1335-1354.
- Guikema, S. D. and Coffelt, J. P. (2008) A flexible count data regression model for risk analysis. *Risk Analysis*, **28**(1), 213-223.
- Lord, D., Guikema, S. D., and Geedipally, S. R. (2008) Application of the Conway-Maxwell-Poisson generalized linear model for analyzing motor vehicle crashes. *Accident Analysis & Prevention*, **40**(3), 1123-1134.
- Kutner, M.H., Nachtsheim, C.J., and Neter, J. (2003) *Applied Linear Regression Models*, 4th edition. McGraw-Hill.
- McCullagh, P. and Nelder, J. A. (1997) *Generalized Linear Models*, 2nd edition. Chapman & Hall/CRC.
- Puig, P. and Valero, J. (2006) Count Data Distributions: Some Characterizations with Applications, *Journal of the American Statistical Association*, **101** (473), 332-340.
- Sellers, K. F., and Shmueli, G. (2008) A Flexible Regression Model for Count Data. Robert H. Smith School Research Paper No. RHS 06-061. Available at SSRN: <http://ssrn.com/abstract=1127359>
- Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., and Boatwright, P. (2005) A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution. *Applied Statistics*, **54**, 127-142.

Smoothing two-dimensional mortality rates with a given percentage of smoothness

Eliud Silva¹, Víctor M. Guerrero² and Daniel Peña¹

¹ Departamento de Estadística, Universidad Carlos III de Madrid, C/ Madrid, 126 - 28903 Getafe (Madrid), Spain.

² Communicating author, Departamento de Estadística, Instituto Tecnológico Autónomo de México (ITAM), 01080 México, D. F., México. E-mail: guerrero@itam.mx.

Abstract: We present a method that allows the analyst to decide at the outset a desired percentage of smoothness in the context of smoothing a two-dimensional mortality table. Thus, we obtain comparable mortality trends for different data sets, in terms of smoothness in the dimension of age, year, or both. The main idea is to use a scalar index that relates the smoothing parameter of a P-spline to: (i) a desired percentage of smoothness and (ii) the sample size available. Some theoretical results that lend support to the use of this index are established and illustrative numerical examples are shown to appreciate the results that can be obtained in practical applications.

Keywords: comparability; index of smoothness; Generalized Least Squares; P-splines; smoothness parameter.

1 Introduction

When analyzing mortality rates, smoothed estimates are of paramount importance to make strategic decisions and planning in population councils or insurance companies. The literature presents several methodological proposals for the analysis, estimation and prediction of mortality rates. In the two-dimensional context we emphasize the work of Currie et al. (2004) who extended the one-dimensional penalized B-splines idea of Eilers and Marx (1996). However, none of the existing works has considered the possibility of controlling the smoothness achieved by the estimates, to allow for valid comparisons of mortality trends in the dimension of age and/or the dimension of year. This is the main objective of this work.

The traditional smoothing approach makes use of a smoothing constant, λ , selected with the aid of an automatic criterion like AIC. Some drawbacks of using automatic criteria can be found in Hastie and Tibshirani (1990). Such an approach is easy to apply, but we have no control on the smoothness achieved. From a purely descriptive point of view we suggest, at least, to measure the amount of smoothness attained by using a particular value

of λ . We go one step further, because our proposal is to fix in advance a desired amount of smoothness for all the mortality curves to be smoothed. In that case, the λ values might be different (depending on the sample sizes) but all the curves will share the same degree of smoothness.

2 Two-dimensional smoothing

We use the vectors of deaths occurred $\mathbf{d} = (d_{11}, \dots, d_{m1}, \dots, d_{1n}, \dots, d_{mn})'$, forces of mortality $\boldsymbol{\mu} = (\mu_{11}, \dots, \mu_{m1}, \dots, \mu_{1n}, \dots, \mu_{mn})'$ and exposures to the risk of dying $\mathbf{e} = (e_{11}, \dots, e_{m1}, \dots, e_{1n}, \dots, e_{mn})'$ for ages 1 to n , and years 1 to m . We assume that d_{im} is a realization of a Poisson process with mean $e_{ij}\mu_{ij}$, for $i = 1, \dots, m$ and $j = 1, \dots, n$. Then, we look for a smoothed estimate of the vector $\mathbf{Y} = (Y_{11}, \dots, Y_{m1}, \dots, Y_{1n}, \dots, Y_{mn})'$, where $Y_{ij} = \log(d_{ij}/e_{ij})$ is the crude force of mortality. Let $B_a = B(x_a)$ be an $m \times m_a$ regression matrix of B-splines based on the explanatory variable x_a for ages. Also, let $B_y = B(x_y)$ be an $n \times n_y$ matrix of B-splines based on x_y for years. Then, $B = B_y \otimes B_a$ is associated to an mn vector $\boldsymbol{\alpha}$ of regression coefficients to be estimated.

The smoothness restrictions in the age and year dimensions are

$$\sum_{j=-3}^{n-1} \sum_{k=-1}^{m-1} (\alpha_{k,j} - 2\alpha_{k-1,j} + \alpha_{k-2,j})^2 = \boldsymbol{\alpha}'(I_n \otimes D'_a D_a)\boldsymbol{\alpha},$$

and

$$\sum_{k=-3}^{m-1} \sum_{j=-1}^{n-1} (\alpha_{k,j} - 2\alpha_{k,j-1} + \alpha_{k,j-2})^2 = \boldsymbol{\alpha}'(D'_y D_y \otimes I_m)\boldsymbol{\alpha}$$

where D_a and D_y are second difference matrices on columns (age) and rows (year), respectively. Then, the smoothing problem can be expressed as

$$\min_{\boldsymbol{\alpha}} [(\mathbf{Y} - B\boldsymbol{\alpha})'(\mathbf{Y} - B\boldsymbol{\alpha}) + \lambda_a \boldsymbol{\alpha}'(I_n \otimes D'_a D_a)\boldsymbol{\alpha} + \lambda_y \boldsymbol{\alpha}'(D'_y D_y \otimes I_m)\boldsymbol{\alpha}].$$

For theoretical purposes we employ two nonsingular $m \times m$ and $n \times n$ natural-spline bases, N_a for age and N_y for years. Thus, the solution makes use of the $mn \times mn$ nonsingular natural-spline basis $N_{ay} = N_a \otimes N_y$ and a transformed version of $\boldsymbol{\alpha}$, corresponding to the new basis, that is

$$\widehat{\mathbf{Y}}_{(\boldsymbol{\alpha})} = (I_{mn} + \lambda_a K'_a K_a + \lambda_y K'_y K_y)\mathbf{Y}$$

with $K_a = (I_n \otimes D_a)N_{ay}^{-1}$ and $K_y = (D_y \otimes I_m)N_{ay}^{-1}$. This is the result obtained by Currie et al. (2004).

3 Proposed solution

The unobserved components model we use is given by

$$\mathbf{Y} = \mathbf{Y}_{(\alpha)} + \mathbf{\Psi}$$

where $\mathbf{Y}_{(\alpha)}$ is a vector of smoothed forces of mortality and $\mathbf{\Psi}$ is a vector of random errors with $E(\mathbf{\Psi}) = \mathbf{0}$ and $Var(\mathbf{\Psi}) = \sigma_{\mathbf{\Psi}}^2 I_{mn}$. We complement it by inducing smoothness in the dimensions of age and year, as follows

$$K_a \mathbf{Y}_{(\alpha)} = \mathbf{\Theta}$$

with $E(\mathbf{\Theta}) = \mathbf{0}$ and $Var(\mathbf{\Theta}) = \sigma_{\mathbf{\Theta}}^2 I_{(m-2)n}$,

$$K_y \mathbf{Y}_{(\alpha)} = \mathbf{\Phi}$$

with $E(\mathbf{\Phi}) = \mathbf{0}$ and $Var(\mathbf{\Phi}) = \sigma_{\mathbf{\Phi}}^2 I_{(n-2)m}$. We also assume $E(\mathbf{\Theta}\mathbf{\Psi}') = 0$ and $E(\mathbf{\Phi}\mathbf{\Psi}') = 0$. Then Generalized Least Squares produces the solution

$$\hat{\mathbf{Y}}_{(\alpha)} = (I_{mn} + \lambda_a K_a' K_a + \lambda_y K_y' K_y) \mathbf{Y}$$

with $\lambda_a = \sigma_{\mathbf{\Psi}}^2 / \sigma_{\mathbf{\Theta}}^2$ and $\lambda_y = \sigma_{\mathbf{\Psi}}^2 / \sigma_{\mathbf{\Phi}}^2$. This solution is identical to that obtained by Currie et al. (2004) using the traditional approach. The resulting Mean Square Error matrix is given by

$$\Gamma = Var(\hat{\mathbf{Y}}_{(\alpha)}) = (\sigma_{\mathbf{\Psi}}^{-2} I_{mn} + \sigma_{\mathbf{\Theta}}^{-2} K_a' K_a + \sigma_{\mathbf{\Phi}}^{-2} K_y' K_y)^{-1},$$

so that its inverse Γ^{-1} is the sum of three precision matrices, $P = \sigma_{\mathbf{\Psi}}^{-2} I_{mn}$, $Q_a = \sigma_{\mathbf{\Theta}}^{-2} K_a' K_a$ and $Q_y = \sigma_{\mathbf{\Phi}}^{-2} K_y' K_y$ associated to the different elements of the model. As did Guerrero (2008) in a univariate time series context, we make use of this fact to measure precision shares attributable to the smoothness elements of the model and to decide the value of the smoothing parameters involved, λ_a and λ_y , as indicated below.

4 Choosing the smoothing parameters

We first establish the following result: a scalar index that measures the proportion of P in $(P + Q_a + Q_y)^{-1}$ is given by

$$\wedge(P; P + Q_a + Q_y) = tr [P(P + Q_a + Q_y)^{-1}] / mn.$$

This is the unique measure that satisfies the following criteria: (i) adding-up, in the sense that $\wedge(P; P + Q_a + Q_y) + \wedge(Q_a + Q_y; P + Q_a + Q_y) = 1$; (ii) boundedness, since it takes on values between zero and one; (iii) invariance under linear nonsingular transformations of the variable involved; and (iv) linearity.

The two-dimensional index of smoothness is given by

$$S_{ay}(\lambda_a, \lambda_y; m, n) = 1 - \text{tr} [(I_{mn} + \lambda_a K'_a K_a + \lambda_y K'_y K_y)^{-1}] / mn.$$

We can see that this index depends only on the constants λ_a and λ_y , as well as on m and n , since K_a and K_y are fully determined by m and n . Thus, we propose to fix at the outset of the study the desired percentage of smoothness $100S_{ay}(\lambda_a, \lambda_y; m, n)\%$ and find the corresponding smoothing parameters that produce such a value.

The common smoothing procedure makes use of a smoothness parameter λ that results from applying an automatic selection criterion. Then, the trace of the smoother matrix, known as *degrees of freedom* (df) of the model is given by $\text{tr} [(I_m + \lambda K'K)^{-1}]$ in the one-dimensional case and by $df = \text{tr} [(I_{mn} + \lambda_a K'_a K_a + \lambda_y K'_y K_y)^{-1}]$ in the two-dimensional one. Hence, the index proposed here can be considered a reparameterization of the df of the model and our proposal is in line with Hastie and Tibshirani's (1990, p. 52) comment: "it is reasonable to select the value of a smoothing parameter simply by specifying the df of the smooth."

Moreover, we can get the bounds for the smoothness index as follows. Let $\gamma_{a,i}$ and $\gamma_{y,j}$ be the eigenvalues of $K'_a K_a$ and $K'_y K_y$, respectively, then: (i)

$$S_{ay}(0, 0; m, n) \rightarrow 0, \text{ (ii) } S_{ay}(0, \lambda_y; m, n) \rightarrow 1 - \left(\sum_{j=1}^{m-2} \frac{1}{1 + \lambda_y \gamma_{y,j}} + 2 \right) / m, \text{ (iii)}$$

$$S_{ay}(0, \infty; m, n) \rightarrow 1 - 2/n, \text{ (iv) } S_{ay}(\lambda_a, 0; m, n) \rightarrow 1 - \left(\sum_{i=1}^{n-2} \frac{1}{1 + \lambda_a \gamma_{a,i}} + \right.$$

$$\left. 2 \right) / n, \text{ (v) } S_{ay}(\lambda_a, \infty; m, n) \rightarrow 1 - 2 \left(\sum_{i=1}^{n-2} \frac{1}{1 + \lambda_a \gamma_{a,i}} + 2 \right) / mn, \text{ (vi)}$$

$$S_{ay}(\infty, 0; m, n) \rightarrow 1 - 2/m, \text{ (vii) } S_{ay}(\infty, \lambda_y; m, n) \rightarrow 1 - \left(\sum_{j=1}^{m-2} \frac{1}{1 + \lambda_y \gamma_{y,j}} + \right.$$

$$\left. 2 \right) / mn, \text{ (viii) } S_{ay}(\infty, \infty; m, n) \rightarrow 1 - 4/mn.$$

Moreover, the marginal index of smoothness attributable to age is given by

$$S_a(\lambda_a; m) = S_{a\bullet}(\lambda_a, 0; m, 1) = 1 - \text{tr} [(I_m + \lambda_a K'_a K_a)^{-1}] / m,$$

where $K_a = D_a N_y^{-1}$ is an $(m - 2) \times m$ matrix. Similarly, the marginal index of smoothness attributable to years is

$$S_y(\lambda_y; n) = S_{\bullet y}(0, \lambda_y; 1, n) = 1 - \text{tr} [(I_n + \lambda_y K'_y K_y)^{-1}] / n,$$

with $K_y = D_y N_x^{-1}$ an $(n - 2) \times n$ matrix. Some other joint and marginal indices may be defined as well and theoretical results that lend support to their use can be established.

5 Illustrative examples and conclusions

We worked out some empirical examples to illustrate the kind of results that can be obtained in practice with our proposal. The examples consider data on mortality for ages 11-100 and years 1947-1999 from the Continuous Mortality Investigation Bureau (CMIB) of the United Kingdom. These data have been analyzed previously by Currie et al. (2004) and the general background can be found there. First, we illustrate the one-dimensional smoothing situation with two possibilities: a single age and different years or different ages and a single year (see FIGURE 1). Here we see that the uncertainty at both ends of the log-mortality series for year 1955 is much greater than in the middle. This is due to the fact that mortality rates at ages 10-25 and greater than 90 have higher variability than at other ages.

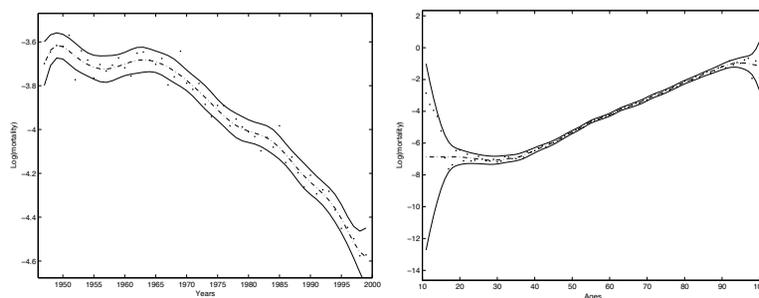


FIGURE 1. Observed and fitted log-mortality with 75% smoothness at age 65 (left panel, $\lambda = 0.1$) and year 1955 (right panel, $\lambda = 461$) with ± 3 standard deviations.

Then, for the two-dimensional case we first replicated the example in Currie et al. (2004, p. 15) for which the smoothness parameters $\lambda_a = 0.6$ and $\lambda_y = 150$ produced the values $AIC = 2306.3$ and $df = 41.2$. We now add that the smoothness attained in this exercise amounts to 75.6% (and the question arises whether that is the smoothness we really want to work with). The other examples in FIGURE 2 allowed us to verify empirically that the log-mortality surfaces may be different for different combinations of parameters, even if they produce the same percentage of joint smoothness (75% in the examples shown). Therefore, we should be aware of the theoretical results here derived about the maximum amount of smoothness that can be attained in practice and the connection between the two-dimensional and the one-dimensional indices.

We conclude that comparability of trends in mortality rates is enhanced by fixing at the outset a desired percentage of smoothness. The index of smoothness we propose allows us to select appropriate smoothing parameters, both in a one-dimensional and two-dimensional settings. Our suggestion arises from the use of GLS that yields identical results to those already established in the literature. Thus, we recommend fixing a desired

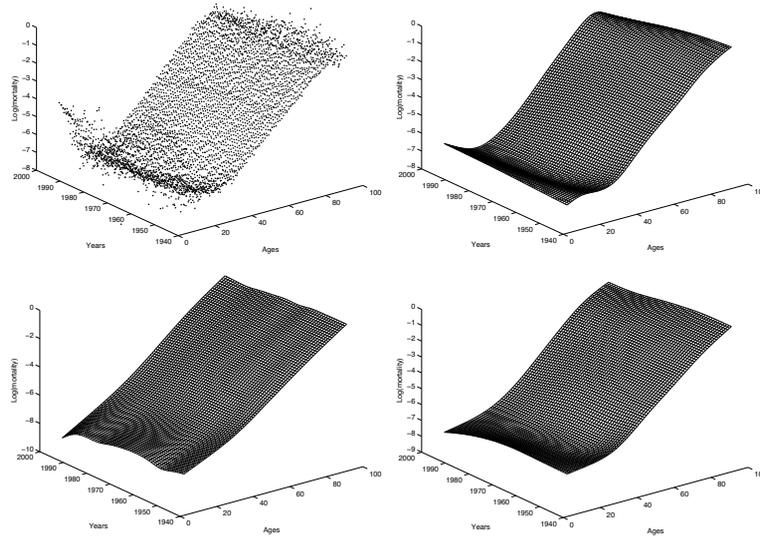


FIGURE 2. Observed log-mortality data (top left) and smoothed by ages 11-100 (top right $\lambda_a = 1$ and $\lambda_y = 2695$), by years (bottom left $\lambda_a = 2440$ and $\lambda_y = 1$) and by both ages and years (bottom right, $\lambda_a = \lambda_y = 227$).

percentage of smoothness and then applying iterative algorithms to find the smoothing parameters that produce such a degree of smoothness.

Acknowledgments: The authors are indebted to M. Durban for making the data on UK mortality and some computing programs available to them. V. M. Guerrero gratefully acknowledges the support provided by Asociación Mexicana de Cultura, A. C. to carry out this work.

References

- Currie, I., Durban, M., and Eilers, P. (2004) Smoothing and forecasting mortality rates. *Statistical Modelling*, **4**, pp. 279–298.
- Eilers, P., and Marx, B. (1996) Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, pp. 89–121.
- Guerrero, V.M. (2008) Estimating Trends with Percentage of Smoothness Chosen by the User. *International Statistical Review*, **76**, Num. 2, pp. 187–202.
- Hastie, T., and Tibshirani R. (1990) *Generalized additive models*. London, Chapman & Hall.

An Additive Penalty Approach to Derivative Estimation of Noisy Data

Andrew Simpkin¹, Paul H. C. Eilers², Jutta Gampe³ and John Newell^{1,4}

¹ School of Mathematics, Statistics and Applied Mathematics, NUI, Galway, Ireland. a.simpkin1@nuigalway.ie

² Department of Biostatistics, Erasmus Medical Centre, Rotterdam, The Netherlands.

³ Max Planck Institute for Demographic Research, Rostock, Germany.

⁴ Biostatistics Unit, Clinical Research Facility, NUI, Galway, Ireland.

Abstract: It is often the case that when analysing data the derivative, or rate of change, of the underlying function describing the observed data is of primary interest. A popular tool for derivative estimation is spline smoothing, with a large number of variants being available. We present a method for derivative estimation which extends the P-spline fitting procedure to include an extra additive penalty term for increased robustness in smoothing. The method is applied to biomedical data and compared with several alternative techniques. A small simulation study is presented to gain further insight about the relative performance of the extended P-spline procedure.

Keywords: Derivative estimation; *P*-splines; Smoothing; Additive penalty; Blood lactate

1 Motivating Example

Blood lactate testing is often used as a measure of endurance in elite athletes. Several features of an individual lactate curve have been considered to be good predictors of endurance performance to track changes in fitness when measured over time. Typically these features, or endurance markers, are used to monitor changes in aerobic fitness, set training regimes and predict endurance performance. However, determination of these markers can be problematic. Newell et al. (2005, 2006) suggest the workload corresponding to the maximum second derivative of the lactate curve (D2LMax) as one such marker. Figure 1 shows a single athletes blood lactate measured at 10 incremental workloads on a treadmill, in addition to the estimated smooth lactate function using a *P*-spline.

2 Smoothing and Derivative Estimation

Eilers & Marx (1996) introduced an alternative penalisation to that used in previous spline smoothing methods (namely the integral of the squared

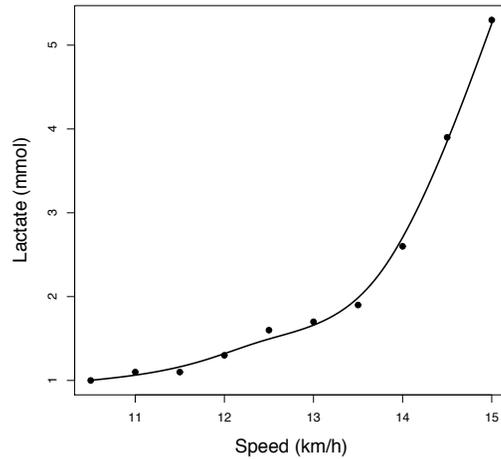


FIGURE 1. Blood lactate data for one athlete with P -spline smoothed curve \hat{l}

second derivative of the fit); they contended that if consecutive coefficients did not differ by much, then the resulting curve would be smooth. Let $B_j(x; q)$ denote the value at x of the j th, $j = 1, \dots, m$, B -spline of degree q . Then a spline f can be written as a linear combination of B -splines

$$f(x) = \sum_{j=1}^m \alpha_j B_j(x; q).$$

Let $\Delta\alpha_j = \alpha_j - \alpha_{j-1}$ where α_j, α_{j-1} are adjacent coefficients for $j = 2, \dots, m$. Penalising $\Delta\alpha_j$ controls their difference and leads to a smooth curve. Moreover, corresponding to the second derivative as a measure of curvature, one penalises second differences $\Delta^2\alpha_j = \alpha_j - 2\alpha_{j-1} + \alpha_{j-2}$. Then f is found by minimising

$$\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \sum_{j=k+1}^m (\Delta^k \alpha_j)^2 \quad (1)$$

where the parameter, λ , controls the degree of smoothing and k is the order of difference which is chosen to be penalised.

Generally derivative estimates are obtained as a by-product of a smooth of the data. In other words, assuming that the data points $(x_i, y_i), i = 1, \dots, n$ originate independently from the model $y_i \sim N(f(x_i), \sigma^2)$ with f being a

smooth function, the first derivative of f is found by taking the derivative of \hat{f} (e.g. for B -splines using the de Boor (2001) formulation). This simple approach comes from the assumption that the optimum derivative estimate is obtained as the derivative of the optimum regression estimate. Whether this is true remains to be seen. For example, the derivatives of data tend to be more sensitive to outliers and as such need a more robust level of smoothing than estimates of f .

3 Problems with Current Techniques

Extensive simulations were carried out in **R** to examine the capabilities of spline smoothing methods for derivative estimation. Many functions including exponential, trigonometric and logarithmic mixed with polynomials in x were studied across a variety of scenarios. One such example is given in Figure 2, which displays first and second derivative estimates of $f(x) = \exp(2x^3) + 3x^4$ with $x_i = U(0, 1)$, $n = 100$ and some small Gaussian error, $\epsilon_i \sim N(0, (0.25(\max[f(x)] - \min[f(x)]))^2)$, added.

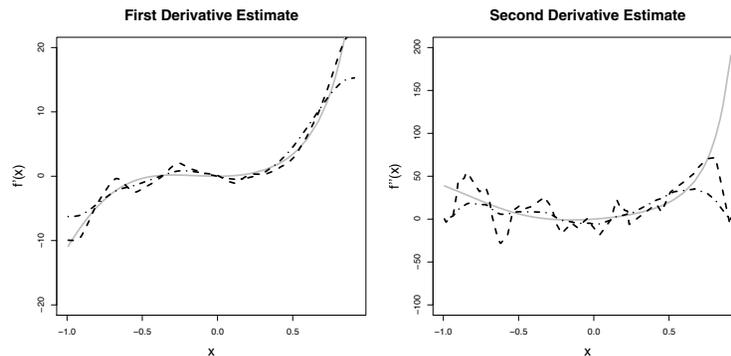


FIGURE 2. Derivative estimates of $\exp(2x^3) + 3x^4$ (solid grey line) using routines `smooth.spline` (dashed line) and `D1D2` (dash-dotted line)

The dashed and dash-dotted curves in Figure 2 represent the fitted values from two derivative estimation routines in **R**. Each routine obtains derivative estimates as a by-product of a spline smoothing fit. As can be seen `smooth.spline` tends to undersmooth (too many ‘wiggles’) whilst both `smooth.spline` and `D1D2` struggle to pick off aspects of the second derivative occurring in the tails (boundary effects). If one is to tune the smoothing parameter to attempt to correct for one aspect, the problem will shift to the other. This idea of sacrificing ‘wiggles’ for a lack of adequately descriptive behaviour is persistent among the several different functions which were

simulated. These deficiencies become more evident the higher the derivative. This can be seen using mean squared error of the derivative

$$RMSED = \sqrt{\frac{1}{n} \sum_{i=1}^n [f^{(l)}(x_i) - \hat{f}^{(l)}(x_i)]^2}$$

as a measure of goodness of fit.

4 Additive Penalty Approach

The methods of Aldrin (2004) and Belitz & Lang (2008) include an additive penalty structure in a P -spline model for increased sensitivity in smoothing a function of the data, e.g. find f which minimises

$$\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda_1 \sum_{j=k+1}^m (\Delta^k \alpha_j)^2 + \lambda_2 \sum_{j=k+2}^m (\Delta^{k+1} \alpha_j)^2. \quad (2)$$

We extend this approach to handle derivative estimation. The extra penalty term allows for increased flexibility which is often required for derivative estimation. Using the additional smoothing term, we can focus extra smoothing in areas which display undersmoothing, while still making sure other aspects are accurately described.

4.1 Smoothing Parameter Selection

An interesting problem raised in the additive penalty model is the selection of multiple smoothing parameters. A simultaneous two dimensional grid search may seem like the natural choice, however, Aldrin (1997) shows that a sequential approach is to be preferred. This sequential approach itself leaves some ambiguity, since one must choose which smoothing parameter to optimise first. Both possibilities to this regard are demonstrated in simulations below.

4.2 Simulation Study

In order to test the performance of the method, four functions (displayed in Figure 3) were chosen to mirror differing scenarios where derivative estimation is applicable. We set $x \sim U[-1, 1]$, $n = 50$ and $\epsilon \sim N(0, 0.5^2)$. The methods under investigation were (1) with $k = 2$, (2) with $k = 2$ and λ_1 chosen first (AP23) and (2) with $k = 2$ and λ_2 chosen first (AP32). In each simulation, the RMSED was recorded for each method, derivative and function. There was evidence that the median RMSED was lower in the additive penalty fits than in the P -spline fit (Kruskal-Wallis $p < 0.0001$) for both the first and second derivative. Generally speaking the AP23 fit

performs better than the AP32 fit and thus in Section 4.3 AP23 will be applied to the lactate data.

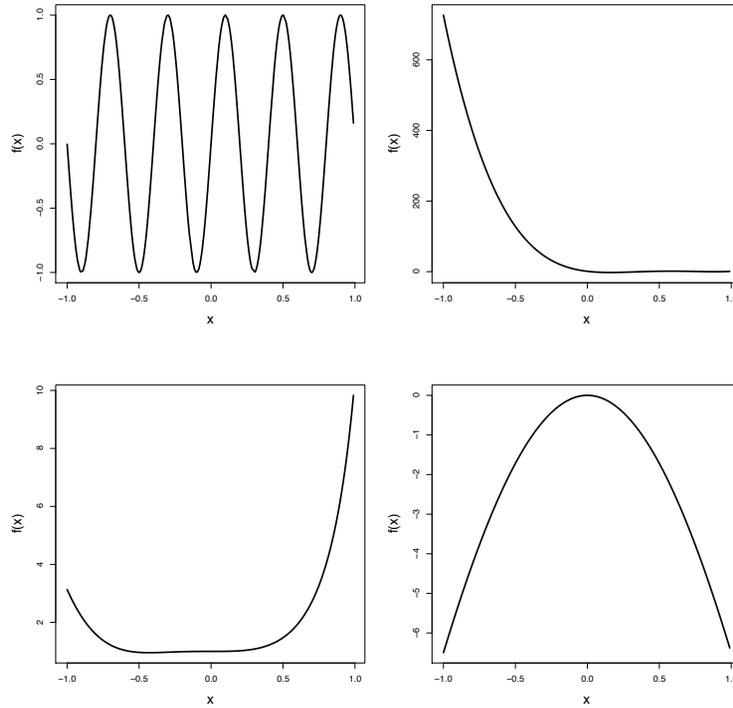


FIGURE 3. $\sin(5\pi x)$ (top left), $1 - 48x + 218x^2 - 315x^3 + 145x^4$ (top right), $\exp(2x^3) + 3x^4$ (bottom left) and $-7x^2 + x^4/2$ (bottom right)

The boxplots in Figure 4 show how the AP23 and AP32 methods display lower median RMSED but also reveal more variability than the P -spline fit. This may be due to the increased variability in selecting two smoothing parameters. Varying $n = 20$ and $n = 100$ the AP methods performed even better than in the $n = 50$ case. Varying $\epsilon \sim N(0, 0.2^2)$ had the same effect whereas $\epsilon \sim N(0, 1)$ decreased the improvement over the P -spline fit. In Figure 5, difference curves show how the additive penalty improves on the P -spline fit, especially in the second derivative estimate.

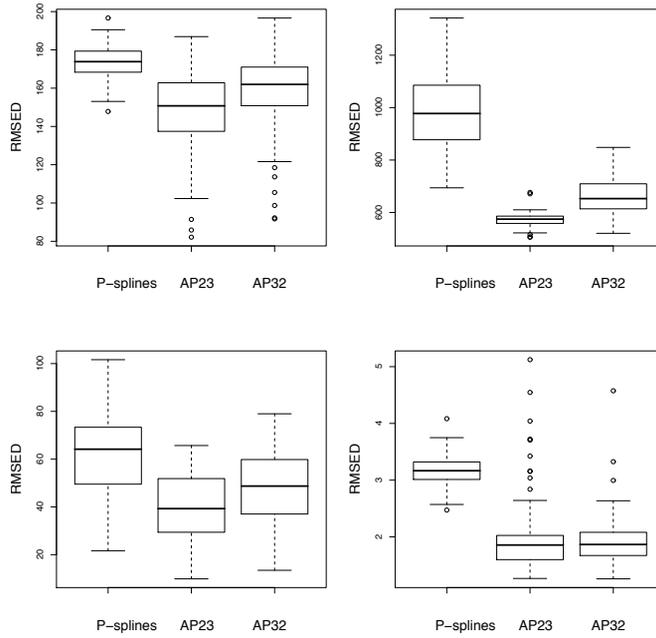


FIGURE 4. RMSED of second derivative estimates for the three methods and four functions panelled as in Figure 3. AP23 uses two penalises second and third order differences, choosing the penalty on the second order penalties first

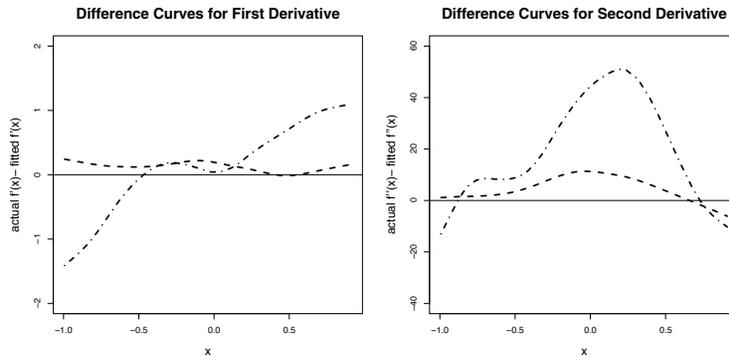


FIGURE 5. Curves of $f'(x) - \hat{f}'(x)$ (left) and $f''(x) - \hat{f}''(x)$ (right) for $\exp(2x^3) + 3x^4$ using AP23 method (dashed line) and P-splines (dash-dotted line)

4.3 Application in Lactate Data

When applying derivative estimation techniques to real data, one needs to make a certain leap of faith since there is no method to test for goodness of fit. From the simulations above it appears that the AP23 method is a better tool to obtain second derivative estimates in practice. For instance, in Fig-

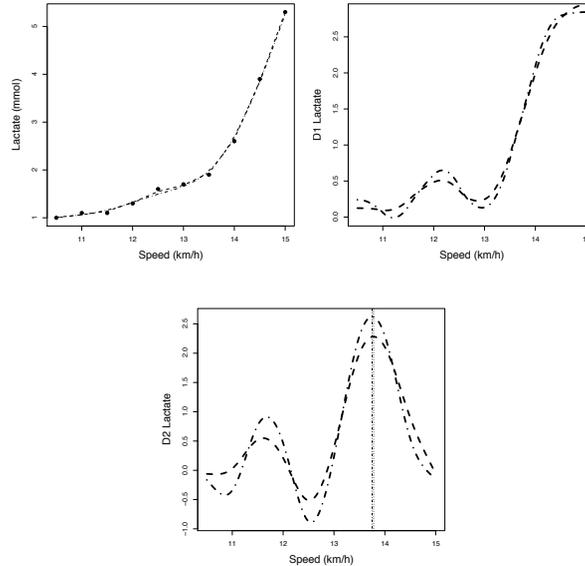


FIGURE 6. AP23 (dashed curve) and P -splines (dash-dotted curve) applied to an individual's lactate data

ure 6 the AP23 and P -spline methods are applied to the individual lactate curve introduced in Section 1. We are interested in picking off the speed at which the mode of the second derivative estimate occurs (D2LMax), using AP23 it may be possible to improve upon these estimates. Both methods choose similar values for D2LMax, however, if we were interested in measuring this maximum, the methods would give different estimates. The P -spline fit is more aggressive than the AP23 method in capturing the derivatives, which is to be expected given the extra smoothing introduced by the additive penalty term.

5 Discussion

Simulations give evidence that the additive penalty method in (2) improves upon modern spline smoothing methods in terms of derivative estimation

in the scenarios considered in this paper. Points of discussion include the extra variability introduced by the additional penalty and how this effects the accuracy of derivative estimates. Further research will evaluate analytically the process of adding extra penalty terms. Whether incorporating some form of adaptive smoothing technique would be beneficial to derivative estimation will also be investigated. Durbán et al. (2004) extend the P -spline model to smooth two dimensional mortality data. Derivative estimation is useful in this application and thus the additive penalty proposed here may be applied in future work.

Acknowledgments: The first author is grateful to the IRCSET for their continued funding. Collaboration of the authors was supported by the Max Planck Institute for Demographic Research and benefitted from fruitful discussions with Giancarlo Camarda. The first author would also like to thank Professor John Hinde for his ongoing support.

References

- Aldrin, M. (1997). Length-modified ridge regression. *Computational Statistics and Data Analysis*. **25**, 377-398.
- Aldrin, M. (2004). Improved predictions penalizing both slope and curvature in additive models. *Computational Statistics and Data Analysis*. **50**, 2672-84.
- Belitz, C. and Lang, S. (2008). Complex additive penalties for generalized structured additive regression. In: *Proceedings of the 23rd IWSM*. 115-120, Utrecht.
- de Boor, C. (2001). *A Practical Guide to Splines*, Revised Edition. New York: Springer.
- Durbán, M., Currie, I. and Eilers, P. H. C. (2002). Using P-splines to smooth two-dimensional Poisson data. In: *Proceedings of the 17th IWSM*. 2072-14, Chania.
- Eilers, P. H. C. and Marx, B. (1996). Flexible Smoothing with B-Splines and Penalties. *Statistical Science*. **11**, 89-121.
- Newell, J., Einbeck, J., Madden, N., McMillan, K. (2005) Model free endurance markers based on the second derivative of blood lactate curves. In: *Proceedings of the 20th IWSM*. 357-364, Sydney.
- Newell, J., McMillan, K., Grant, S., McCabe, G. (2006) Using functional data analysis to summarise and interpret lactate curves. *Computers in Biology and Medicine*. **36**, 262-275.

Investigating smoking behaviour as a risk factor for stroke-free life expectancy

Ardo van den Hout¹ and Fiona E. Matthews¹

¹ MRC Biostatistics Unit, Institute of Public Health, Cambridge University.
Robinson Way, Cambridge CB2 0SR, U.K. E-mail: ardo.vandenhout@mrc-bsu.cam.ac.uk.

Abstract: Stroke-free life expectancy is the expected remaining number of years spent free of stroke. When stroke history is described by a progressive three-state illness-death model where state one denotes the absence of a history of stroke, stroke-free life expectancy can be estimated using established multi-state models and the effect of a risk factor such as smoking behaviour on transition intensities can be investigated. This however does not provide the direct effect of the risk factor on stroke-free life expectancy. A method is presented which makes it possible to investigate this direct effect given interval-censored longitudinal data. Multiple imputation is used to estimate unobserved transition times and the Tobit model is applied to estimate the direct effect.

Keywords: Life expectancy, Multi-state model; Multiple imputation; Survival.

1 Introduction

We consider data where history of stroke can be described by a three-state illness-death model. State 1 is the stroke-free state, state 2 denotes a history of one or more strokes, and state 3 is the death state. Given a specified individual and a point in time, stroke-free life expectancy (LE) is the expected number of years this individual will spent in state 1. We are interested in smoking behaviour as a risk factor for stroke-free LE.

In the data, states are right-censored at the end of the follow-up, all death times during follow-up are recorded, and transitions from state 1 to state 2 are interval-censored. In this setting, it is possible to estimate a multi-state model and to use the model parameters to estimate LE. Covariates can be related to transition intensities and can be taken into account when estimating LE. The disadvantage of this approach is that the covariates effects are with respect to the transition intensities and are not directly linked to LE. This paper proposes a method to quantify the effect of covariates on LE. The basic idea is to impute exact transition times for the transitions from state 1 to state 2 during follow-up. Given exact transition times, the right-censored survival times in state 1 are available and standard survival models can be applied to investigate covariate effects and to estimate LE.

This paper investigates the applicability of the Tobit model. To take into account the uncertainty of the imputation, we use multiple imputation (MI) techniques as introduced by Rubin (1987).

Scheike and Zhang (2007) show how to estimate regression effects on (functions of) transition probabilities. Since stroke-free can be estimated by a function of transition probabilities, Scheike and Zhang have the same aim as we. In addition to dealing with interval-censored data, we consider our method to be more simple: it does not involve new theory and there is no need to estimate the censoring distribution for each individual. Tunes-da-Silva and Klein (2009) investigate methods to estimate regression effects on mean quality-adjusted lifetime. This work is very recent and we are still looking into the similarities and the differences between their work and ours. Important differences seem to be that Tunes-da-Silva and Klein (2009) do not discuss interval-censored data and that they provide a regression model for *restricted* mean quality of life - a concept that does not cover life expectancy.

2 Data

The Medical Research Council Cognitive Function and Ageing Study (MRC CFAS, www.cfes.ac.uk) is a population based longitudinal study of cognition and health in the older population of England and Wales. We use a subset of the data of men in rural Cambridgeshire. All men were aged 65 years and above at baseline and all deaths up to the end of 2005 have been included. We consider the 1041 men that were not severe cognitively impaired at baseline 1991 so as to exclude bias in the baseline status. Time between interviews varies between and within men and the number of interviews is not fixed. The last observed state at the end of December 2005 is either death or censored. Given the definition of state 2, there are no transitions from state 2 back to state 1.

3 Models

There are three stages in our approach: **(I)** the fitting of a time-continuous three-state model, **(II)** imputation of exact transition times, and **(III)** fitting the Tobit model.

Ad **I**. There is no new theory here - the basic ideas can be found in Kalbfleisch and Lawless (1985). Extended versions that can be used to estimate LE are discussed in Van den Hout and Matthews (2008). The model for the CFAS data approximates the time-dependency of the transition intensities by using a piecewise-constant intensity model with age as a time-dependent covariate. This model can be fitted using the R package

msm (Jackson *et al.*, 2003). Stroke-free LE is given by

$$\int_0^\infty \mathbb{P}(X_t = 1 \mid X_0 = 1, \mathcal{Z}) dt,$$

where X_t denote the state at time t and \mathcal{Z} is the covariate history that is assumed to be deterministic. The integrand $\mathbb{P}(X_t = 1 \mid X_0 = 1, \mathcal{Z})$ can be derived from the fitted three-state model.

Ad (II). The proposed method is as follows.

1. From the fitted three-state illness-death model, extract the estimated maximum likelihood parameter vector and its estimated variance-covariance matrix.
2. Draw M parameter vectors from the assumed multivariate normal distribution of the maximum likelihood estimator.
3. Impute M exact transition times for individually observed intervals that should or might contain a transition from state 1 to state 2.

In step 3, the transition time to state 2 is imputed if the interval-censored observation is $1 \rightarrow 2$. A transition time might be imputed if the interval-censored observation is $1 \rightarrow 3$ or $1 \rightarrow \text{censored state}$. In these cases, firstly, it is simulated whether or not state 2 was visited, and, secondly - if state 2 was visited - the time at which the state was entered is imputed.

Given a simulated parameter vector, we can derive the transition intensities for any given time interval. Consider an observed interval $(t, u]$. This interval is partitioned using a small chosen $\epsilon > 0$ to approximate continuous time. Using the Markov assumption, for $(t, u]$ partitioned into $t_1, t_2, \dots, t_{j-1}, t_j$ equal to $t, t_1 + \epsilon, t_1 + 2\epsilon, \dots, u$, it follows that

$$\begin{aligned} \mathbb{P}(X_{t_j} = 2 \mid X_{t_1} = \dots = X_{t_{j-1}} = 1, X_{t_j} = 2) \\ = \frac{\mathbb{P}(X_{t_j} = 2 \mid X_{t_j} = 2) \mathbb{P}(X_{t_j} = 2 \mid X_{j-1} = 1)}{\mathbb{P}(X_{t_j} = 2 \mid X_{t_{j-1}} = 1)}. \end{aligned}$$

MI of the (ϵ -approximated) exact transition time for $(t, u]$ can now be implemented using a series of Bernoulli distributions.

In case the interval-censored observation is $1 \rightarrow 3$, we first need to estimate $\mathbb{P}(X_{u-} = 2 \mid X_t)$, where X_{u-} denotes the state just before death at time u . Secondly we simulate from the assumed Bernoulli distribution with parameter $\mathbb{P}(X_{u-} = 2 \mid X_t)$ whether or not state 2 was visited. Thirdly - if state 2 was visited - we use the procedure above to impute an exact transition time for the trajectory $1 \rightarrow 2$. We approximate $\mathbb{P}(X_{u-} = 2 \mid X_t)$ by estimating $\mathbb{P}(X_u = 2 \mid X_t)$ which is readily provided by the fitted three-state model. For the interval-censored observation $1 \rightarrow \text{censored state}$ we can use the same approach using $\mathbb{P}(X_u = 2 \mid X_t)$ directly to simulate the latent state at time u .

TABLE 1. Stroke-free life expectancies for men with less than 10 years of education. Estimated standard errors in parentheses.

Age at baseline	Three-state model	Tobit model
<i>Non-smokers and ex-smokers</i>		
65	13.68 (0.40)	13.31 (1.01)
75	9.07 (0.31)	9.22 (0.76)
85	5.68 (0.31)	5.12 (1.14)
<i>Current-smokers</i>		
65	10.48 (0.28)	11.40 (0.99)
75	6.69 (0.20)	7.30 (0.74)
85	4.05 (0.21)	3.21 (1.13)

Ad (III). Fit M times a time-to-event model to the survival times in state 1 and investigate the effect of the risk factor of interest. Because times are multiple imputed, Rubin's MI rules are used to combine the results. The Tobit model is a linear regression model which takes possible right-censoring into account. It can be fitted by using the R package `survival`.

4 Analysis

Covariates in both the three-state model and in the Tobit model are age, education (0 = less than 10 years of education, 1 = 10 years or more) and smoking behaviour (0 = non-smoker or ex-smoker, 1 = current smoker). Time is in months. We estimated stroke-free LE using both models to see whether results for the Tobit model with the imputed data are similar to the results of the three-state model. We used $M = 10$ for the imputation. Covariate regression effects and LE is estimated using Rubin's rule for MI. For the Tobit model, the regression effects are given by

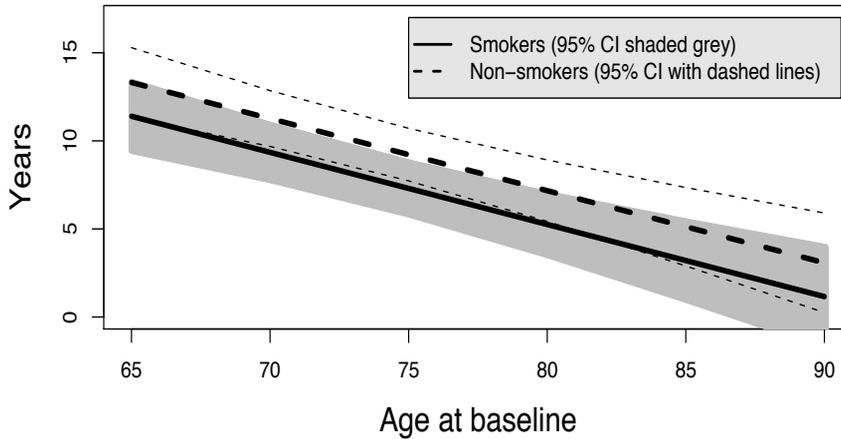
$$\begin{aligned}\widehat{\beta}_{Age} &= -4.91 (0.37), & \widehat{\beta}_{Education} &= 15.22 (5.21), & \text{and} \\ \widehat{\beta}_{Smoking} &= -23.03 (4.64),\end{aligned}$$

where estimated standard errors are in parentheses. Table 1 with the LEs shows that results are indeed similar although there is room for improvement.

The direct effect of smoking is given by the covariate effect for smoking behaviour in the Tobit model: -23.03. Since the Tobit model is a linear regression model, the interpretation of this effect is straightforward: being a current smoker reduces your stroke-free LE with 23.03 months compared to a non-smoker of the same age from the same education group.

For age at baseline, Figure 1 shows estimated stroke-free life expectancies for men with less than 10 years of education.

FIGURE 1. Stroke-free life expectancies for men with less than ten years of education. Results for the Tobit model, with 95% confidence intervals (CIs).



5 Conclusion

Other models than the Tobit model are of course possible - for example the Weibull or the model with a loglogistic distribution. However for the current data the Tobit model performs best when estimated LEs are compared. We consider this research work in progress. For now we have chosen a parametric survival model in order to extrapolate for the estimation of LE.

References

- 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.

Scheike, T.H., and Zhang, M.-J. (2007). Direct modelling of regression effects for transition probabilities in multistate models. *Scandinavian Journal of Statistics*, **34**, 17-32.

Tunes da Silva, G., and Klein, J.P. (2009). Regression analysis of mean quality-adjusted survival time based on pseudo-observations, *Statistics in Medicine*, DOI: 10.1002/sim.3529.

Van den Hout, A., and Matthews, F.E. (2008). Multi-state analysis of cognitive ability data: a piecewise-constant model and a Weibull model, *Statistics in Medicine* **27**, 5440-5455.

Birnbaum-Saunders random intercept models with censored data

Cristian Villegas¹, Gilberto A. Paula² and Víctor Leiva³

¹ Dept of Statistics, Universidade de São Paulo, Brazil, e-mail:clobos@ime.usp.br

² Dept of Statistics, Universidade de São Paulo, Brazil, e-mail:giapaula@ime.usp.br

³ Dept of Statistics, Universidad de Valparaíso, Chile, e-mail:victor.leiva@uv.cl

Abstract: In this work, we introduce Birnbaum-Saunders random intercept models with censored data. Specifically, we estimate their parameters by using the Gauss-Hermite quadrature approximation, carry out a residual analysis for these models, and discuss an example with real data as illustration.

Keywords: GH quadrature; Random effect models; Sinh-normal distribution.

1 Introduction

The Birnbaum-Saunders (BS) distribution (Birnbaum-Saunders, 1969) is based on a physical argument of cumulative damage that produces fatigue in the materials. This distribution was derived from a model that shows the total time that passes until some type of cumulative damage surpasses a threshold value and causes the material specimen to fail. Various authors have developed different aspects related to BS models and applied them to several fields; however, in general, fixed effects have been assumed; see, for example, Leiva et al. (2007), Barros et al. (2008), and references therein. The aim of this work is to introduce BS random intercept models with censored data, estimate their parameters by using the Gauss-Hermite (GH) quadrature approximation, and carry out a residual analysis for these models. The proposed methodology is applied to a real data set.

2 The Birnbaum-Saunders distribution

The BS distribution is defined in terms of the normal model by means of the r.v. $T = \beta[\alpha Z/2 + \sqrt{[\alpha Z/2]^2 + 1}]^2$, where $Z \sim N(0, 1)$, $\alpha > 0$ is the shape parameter, and $\beta > 0$ is the scale parameter and the median. This

¹Address correspondence to Cristian Villegas. Instituto de Matemática e Estatística, USP - Caixa Postal 66281 (Ag. Cidade de São Paulo), 05311-970 São Paulo - SP - Brasil. E-mail: clobos@ime.usp.br

is denoted by $T \sim \text{BS}(\alpha, \beta)$. The pdf of T is given by

$$f_T(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2\alpha^2} \left[\frac{t}{\beta} + \frac{\beta}{t} - 2\right]\right) \frac{t^{-3/2}[t + \beta]}{2\alpha\sqrt{\beta}}, \quad t > 0. \quad (1)$$

If $T \sim \text{BS}(\alpha, \beta)$, then $Z = [\sqrt{T/\beta} - \sqrt{\beta/T}]/\alpha \sim \text{N}(0, 1)$. Rieck and Nedelman (1991) developed the sinh-normal (SN) distribution, which is characterized by its shape, location and scale parameters, $\alpha > 0$, $\mu \in \mathbb{R}$, and $\sigma > 0$, respectively. Thus, for an r.v. Y , the notation $Y \sim \text{SN}(\alpha, \mu, \sigma)$ is used. The SN distribution is symmetrical around μ , strongly unimodal for $\alpha \leq 2$, and bimodal for $\alpha > 2$. If $T \sim \text{BS}(\alpha, \beta)$, then $Y = \log(T) \sim \text{SN}(\alpha, \mu, \sigma = 2)$, where $\mu = \log(\beta)$.

3 A log-BS random intercept model

Let y_{ij} denote the j th log-outcome measured for the i th cluster (subject), for $i = 1, \dots, n$ and $j = 1, \dots, m_i$. We assume the log-BS random intercept model:

- (i) $Y_{ij}|b_i \stackrel{\text{ind.}}{\sim} \text{log-BS}(\alpha, \mu_{ij})$ and
- (ii) $b_i \stackrel{\text{ind.}}{\sim} \text{N}(0, \varsigma)$,

where $\mu_{ij} = \eta_{ij} + b_i$, $\eta_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta}$, $\mathbf{x}_{ij} = [x_{ij1}, \dots, x_{ijp}]^\top$ contains values of explanatory variables, and $\boldsymbol{\beta} = [\beta_1, \dots, \beta_p]^\top$. Note that $\text{E}[Y_{ij}] = \eta_{ij}$, $\text{Var}[Y_{ij}] = 4\omega(\alpha) + \varsigma$, and $\text{Cov}[Y_{ij}, Y_{ij'}] = \varsigma$, for $j \neq j'$, where $\omega(\alpha)$ may be obtained from the respective moment generating function; see Rieck and Nedelman (1991). Then, the intraclass correlation yields $\text{Corr}[Y_{ij}, Y_{ij'}] = \varsigma/[4\omega(\alpha) + \varsigma]$, for $j \neq j'$. The marginal pdf of $\mathbf{Y} = [\mathbf{Y}_1^\top, \dots, \mathbf{Y}_n^\top]^\top$, with $\mathbf{Y}_i = [Y_{i1}, \dots, Y_{im_i}]^\top$, is given by $f(\mathbf{y}|\alpha, \boldsymbol{\beta}, \varsigma) = \prod_{i=1}^n f_i(\mathbf{y}_i|\alpha, \boldsymbol{\beta}, \varsigma)$, where

$$f_i(\mathbf{y}_i|\alpha, \boldsymbol{\beta}, \varsigma) = \int_{-\infty}^{+\infty} \left[\prod_{j=1}^{m_i} f_{ij}(y_{ij}|b_i, \alpha, \boldsymbol{\beta}) \right] f(b_i|\varsigma) db_i. \quad (2)$$

4 Parameter estimation

The total log-likelihood function for $\boldsymbol{\theta} = [\alpha, \boldsymbol{\beta}^\top, \varsigma]^\top$, denoted by $L(\boldsymbol{\theta})$, with censored data may be approximated by the GH quadrature; see McCullogh and Searle (2001). We find $L(\boldsymbol{\theta}) \approx \sum_{i=1}^n \log(D_i(\boldsymbol{\theta}))$, where

$$D_i(\boldsymbol{\theta}) = \sum_{k=1}^q \frac{v_k}{\sqrt{\pi}} A_{ik}(\boldsymbol{\theta}) B_{ik}(\boldsymbol{\theta}),$$

$$A_{ik}(\boldsymbol{\theta}) = \exp\left(\sum_{j \in D} \left\{ -\frac{1}{2} \log(8\pi) + \log(\xi_{ijk1}) - \frac{1}{2} \xi_{ijk2}^2 \right\}\right),$$

$$B_{ik}(\boldsymbol{\theta}) = \exp\left(\sum_{j \in C} \log(\Phi(-\xi_{ijk2}))\right),$$

$\xi_{ijk1} = [2/\alpha] \cosh(H_{ijk})$, $\xi_{ijk2} = [2/\alpha] \sinh(H_{ijk})$, $\xi_{ijk3} = \xi_{ijk2}/\xi_{ijk1} = \tanh(H_{ijk})$, and $H_{ijk} = [y_{ij} - x_{ij}^\top \beta - \sqrt{2\zeta} s_k]/2$. Here, D and C represent the sets of uncensored and censored observations, respectively, whereas s_k , v_k , and $\Phi(\cdot)$ denote the k th zero quadrature point, the k th weight, and the standard normal cdf, respectively. Similarly to Barros et al. (2008), the BFGS method has been applied for maximizing $L(\theta)$.

5 Residual analysis

To study departures from the assumptions of the error terms of the model given in (i) and (ii), as well as presence of outlying observations, we consider the conditional martingale-type residual expressed as

$$r_{\text{MT}_{ij}} = \text{sign}(r_{\text{M}_{ij}}) \sqrt{-2[r_{\text{M}_{ij}} + \delta_j \log(\delta_j - r_{\text{M}_{ij}})]}, \tag{3}$$

where $\delta_j = 0, 1$ indicates whether the observation is censored or not, respectively, $\hat{\mu}_{ij} = x_{ij}^\top \hat{\beta} + \tilde{b}_i$, \tilde{b}_i is the empirical Bayes estimate, and $r_{\text{M}_{ij}} = \delta_j + \log(\Phi(-[2/\hat{\alpha}] \sinh([y_{ij} - \hat{\mu}_{ij}]/2)))$ is the martingale residual; for more details, see Leiva et al. (2007).

6 Application

We consider a data set described by Weibull regression models in Smith (1991). These data correspond to failure times of a particular kind of fiber (107 Kevlar 49) submitted to different stress levels (measured in MPa). A sample of eight spools was considered and each spool was submitted to different stress levels with replicates. Since the spools can be considered random samples from a population of spools, we assume the following log-BS random intercept model to analyze the fiber’s data:

$$y_{ij} = \beta_1 + \beta_2 x_{ij} + b_i + \epsilon_{ij}, \quad i = 1, \dots, 8, \quad j = 1, \dots, m_i, \tag{4}$$

where y_{ij} denotes the logarithm of the failure time or of the censoring time of the j th fiber subject to the stress level x_{ij} coming from the i th spool, $b_i \stackrel{\text{ind.}}{\sim} N(0, \zeta)$ and $\epsilon_{ij} \stackrel{\text{ind.}}{\sim} \text{log-BS}(\alpha, 0)$, with b_i and ϵ_{ij} also being independent. The parameter estimates based on 30 points of the GH quadrature and their corresponding significance levels are given in Table 1. As we can note, all parameters are significant. In particular, the random effect parameter appears to be highly significant. From Figure 1, we do not observe unusual features so that the assumptions of log-BS error and normal random effect do not seem to be unsuitable.

Acknowledgments: The authors are grateful to CNPq and FAPESP, Brazil and FONDECYT, Chile.

TABLE 1. Parameters estimates of the model given in (i)-(ii) fitted to the fiber's data and their corresponding significance levels.

Parameter	Estimate	Std. error	DF	t-value	p-value
α	1.2223	0.0928	7	13.17	0.000
β_1	33.4988	1.5441	7	21.69	0.000
β_2	-0.9900	0.0529	7	-18.71	0.000
ς	1.4305	0.3924	7	3.65	0.008

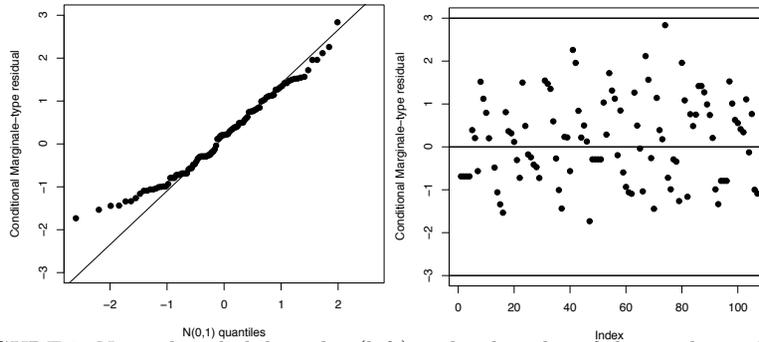


FIGURE 1. Normal probability plot (left) and index plot of the conditional marginale-type residuals (right).

References

- Birnbaum, Z.W., and Saunders, S.C. (1969). A new family of life distributions. *Journal of Applied Probability*, **6**, 319-327.
- Barros, M., Paula, G.A., and Leiva, V. (2008). A new class of survival regression models with heavy-tailed errors: robustness and diagnostics. *Lifetime Data Analysis*, **14**, 316-332.
- Leiva V., Barros M., Paula G.A., and Galea M. (2007). Influence diagnostics in log-Birnbaum-Saunders regression models with censored data. *Computational Statistics and Data Analysis*, **51**, 5694-5707.
- McCulloch, C., and Searle, S. (2001). *Generalized, Linear and Mixed Models*. Wiley, New York.
- Rieck, J.R., and Nedelman, J.R. (1991). A log-linear model for the Birnbaum-Saunders distribution. *Technometrics*, **33**, 51-60.
- Smith, R.L. (1991). Weibull regression models for reliability data. *Reliability Engineering and System Safety*, **34**, 55-77.

Modelling of covariance structure in constrained marginal models for longitudinal data

Jing Xu¹ and Gilbert MacKenzie¹

¹ Centre of Biostatistics, Department of Mathematics & Statistics, University of Limerick, Ireland. Email: jing.xu@ul.ie; gilbert.mackenzie@ul.ie

Abstract: A data-driven method (Pourahmadi 1999, Pan and MacKenzie 2003) for modelling intra-subject covariance matrix is introduced to constrained marginal models with longitudinal data. A constrained iteratively re-weighted least squares algorithm is presented consequently. Asymptotic properties of the constrained ML estimates, including strong consistency, approximate representation and asymptotic distribution, are given. Real data analysis is conducted to compare the data-driven covariance modelling method with classical menu-selection-based modelling technique under constrained models. In addition, during the analysis, we modify the model given by Pourahmadi (1999) for estimating generalised autoregressive parameters.

Keywords: longitudinal data; inequality constraints; marginal models; covariance modelling; Cholesky decomposition

1 Introduction

In longitudinal studies, constrained problems are often interested by the researchers in many practical fields, especially in biomedicine and clinical trials (Tan *et al*, 2005, Fang *et al* 2006, Pilla *et al* 2006, Cysnerios & Paula 2004, Park *et al* 1998.). In order to take account of intra-subject correlations, so called menu selection or working correlation structure are used extensively in these literatures. However, this conventional approach may not work well in some cases (Wang & Carey 2003, Dobson 2002, Pan & MacKenzie 2007).

Based on the work by Xu & Wang (2008), mean modelling with inequality constraints is joint with a data-driven covariance modelling for constrained marginal models in this paper. We use a modified Cholesky decomposition to decompose the within-subject covariance matrices and then parsimoniously model the within-subject correlation and variation in terms of simple regression models. This method presented in the seminal work by Pourahmadi (1999) has several advantages, namely: it is unique and positive definite, its parameters are unconstrained and have useful statistical interpretations, it can reproduce a wide range of classical, stationary

and non-stationary, covariance structure. Specifically, its real strength is in modelling nonstationary features where variances increase over time, and measurements equidistant in time are not equicorrelated. This is demonstrated by cattle data analysis in Pourahmadi(1999). Constrained maximum likelihood estimation is applied to obtain the estimators for the mean and covariance parameters and the estimators are shown to be consistent and asymptotically normally distributed piecewisely. Analysis of diabetic patient data shows that the proposed approach still works well compared to menu selection method when the data are strongly correlated and have a stationary covariance structure.

2 Constrained Marginal Models and Covariance Modelling

Suppose the i -th individual is observed on m_i occasions for $i = 1, \dots, n$. The vector of responses is denoted by y_i . Assume that y_i arises from the constrained marginal model

$$\begin{aligned} y_i &= X_i\beta + \varepsilon_i \quad \text{for } i = 1, \dots, n \\ \text{s.t. } A\beta &\geq b \end{aligned} \quad (1)$$

where *s.t.* is the abbreviation for "subjected to"; X_i is a known $m_i \times p$ design matrix for the i -th individual; β is a $p \times 1$ vector of unknown coefficients to be estimated; A is a $l \times p$ matrix and $b = (b_1, \dots, b_l)'$ is a $l \times 1$ vector; $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{im_i})'$ are independently distributed as $N(0, \Sigma_i)$ for $i = 1, \dots, n$. The inequality constraint set $\{\beta : A\beta \geq b\}$ is quite general containing order restriction as a special case.

Since the subject-specific covariance matrix Σ_i is positive definite, there exists a unique lower triangular matrix T_i with 1's as main diagonal entries and a unique diagonal matrix D_i with positive diagonal entries such that $T_i\Sigma_iT_i' = D_i$. Let ϕ_{ijk} be the jk th below-diagonal entry of T_i and σ_{ij}^2 be the ij -th diagonal entry of D_i where $1 \leq j \leq m_i$ and $1 \leq i \leq n$. The parameters ϕ_{ijk} and $\varsigma_{ij} \equiv \log\sigma_{ij}^2$ are modelled as $\phi_{ijk} = z'_{ijk}\gamma$ and $\varsigma_{ij} = h'_{ij}\lambda$, an augmented linear model for the new regression parameters of interest γ and λ . Then the loglikelihood function for data y_1, \dots, y_n is given by

$$L = -\frac{1}{2} \sum_{i=1}^n m_i \log(2\pi) - \frac{1}{2} \sum_{i=1}^n \log|T_i^{-1}D_iT_i'| - \frac{1}{2} \sum_{i=1}^n r_i'T_i'D_i^{-1}T_i r_i, \quad (2)$$

where $r_{ij} = y_{ij} - x'_{ij}\beta$ is the j th element of $r_i = y_i - X_i\beta$, the vector of residual, and the matrix X_i has row vectors $x'_{ij}(j = 1, 2, \dots, m_i)$.

3 Constrained Maximum Likelihood Estimation

3.1 Estimation of parameters

Denote the feasible solution set of the constrained model (1) by $S = \{\beta : A\beta \geq b\}$ and the function (2) by $L(\beta, \gamma, \lambda)$. Then the constrained ML estimation problem is

$$\max_{\beta \in S} L(\beta, \gamma, \lambda). \tag{3}$$

Based on the iteratively re-weighted least squares algorithm given by Pan and MacKenzie (2003), we present the following procedure for the constrained estimation problem (3).

Given γ and λ , the constrained regression parameters are determined by

$$\beta = \arg \min_{\beta \in S} \sum_{i=1}^n r_i' \Sigma_i^{-1} r_i. \tag{4}$$

Secondly, given β and λ , the first order estimating equation for γ is

$$U_2(\gamma) = \sum_{i=1}^n Z_i^* D_i^{-1} (r_i - Z_i^* \gamma) = 0, \tag{5}$$

where the matrix Z_i^* , of order $m_i \times (q+1)$, has typical row $z_{ij}^* = \sum_{k=1}^{j-1} r_{ik} z'_{ijk}$. Finally, given β and γ , the estimating equation for λ is

$$U_3(\lambda) = \frac{1}{2} \sum_{i=1}^n H_i' (D_i^{-1} e_i - 1_{m_i}) = 0, \tag{6}$$

where $H_i = (h'_{i1}, h'_{i2}, \dots, h'_{im_i})'$, $e_i = (e_{i1}, e_{i2}, \dots, e_{im_i})'$ with $e_{ij} = (r_{ij} - \hat{r}_{ij})^2$ and $\hat{r}_{ij} = \sum_{k=1}^{j-1} \phi_{ijk} r_{ik}$, are the $m_i \times (d+1)$ matrix of covariates and the $m_i \times 1$ vector of squared fitted residuals, respectively, and 1_{m_i} is the $m_i \times 1$ vector of 1's.

The iteration procedure proceeds within (4)-(6) by initializing at $\Sigma_i = I_{m_i}$ ($i = 1, 2, \dots, n$) and iterating until convergence. We refer to it as a constrained iteratively re-weighted least squares algorithm.

3.2 Asymptotic properties

For the sake of simplicity, we denote all of the unknown parameters by $\theta = (\beta', \gamma', \lambda')'$ and the unknown true values by $\theta_0 = (\beta'_0, \gamma'_0, \lambda'_0)'$. Let $a'_j, j = 1, \dots, l$, be the rows of the matrix A . Define a $(p + q + d + 3)$ -dimension vector $A_j = (a'_j, 0, \dots, 0)'$, for $j = 1, \dots, l$. Then the constrained ML estimation problem (3) becomes

$$\begin{aligned} \max \quad & L(y_1, \dots, y_n; \theta) \\ \text{s.t.} \quad & A'_j \theta \geq b_j, \quad \text{for } j = 1, \dots, l. \end{aligned} \tag{7}$$

The optimization solution or the constrained estimators of problem (7) is denoted by $\hat{\theta}$. Appealing to the Kuhn-Tucker conditions and the first- and second-order conditions of the loglikelihood function, the asymptotic properties of $\hat{\theta}$ including consistency, approximate representation and asymptotic distribution can be established (Theorems 1-3 are omitted here). Briefly speaking, under some necessary regularity conditions, the constrained ML estimators $\hat{\theta} = (\hat{\beta}', \hat{\gamma}', \hat{\lambda}')'$ is strongly consistent for the true value $\theta_0 = (\beta_0', \gamma_0', \lambda_0')'$ and the constrained ML estimator $\hat{\theta}$ has a piecewise asymptotic normal distribution.

4 Analysis of Diabetic Patient Data

We reanalyze in this section the Example 2.1 discussed by Crowder and Hand (1990) on a comparative study among diabetic groups. This set of data is also used by Shin et al. (1996) and Cysneiros and Paula (2004) for inequality hypothesis testing. Originally, there are four patient groups in the data. We only consider the first three groups: control group ($n_1 = 8$), diabetic group without complications ($n_2 = 6$) and diabetic group with hypertension ($n_3 = 7$). For each patient the response was a physical task measured in the times 1, 2, 3, 4, 5, 6, 8 and 10 min. So this data set is balanced but irregular.

Scatterplots of responses against time for these three groups show that three constant means may be appropriate for the data. Let y_{ilj} be the observed physical task for the i th patient of the l th group at the time j . We assume the model

$$\mathbf{y}_{il} = \mathbf{u}_l + \varepsilon_{il} \quad (8)$$

where $\mathbf{u}_l = \mu_l \mathbf{1}_m$, $\mathbf{y}_{il} = (y_{il1}, \dots, y_{ilm})^T$ and $\varepsilon_{il} \sim N(0, \Sigma_i)$ with Σ_i being the same over all the subjects. In addition, it is reasonable to assume the constraints $\mu_1 \geq \mu_2 \geq \mu_3$ for the expected values of the physical task. This assumption of constraint is revealed by the scatterplots and also concluded by Shin et al.(1996) and Cysneiros and Paula (2004).

Furthermore, the sample variances(along the main diagonal in Table 1) and sample correlation matrix (above the main diagonal in Table 1) suggest that a stationary covariance structure with strong correlation seems to be very reasonable for the data as all the responses are measured during ten minutes. Classic structures, such as compound symmetry, AR(1) and ARMA(1,1), may be good models for the data. Actually, compound symmetry and AR(1) structure are considered in Shin et al.(1996) and Cysneiros and Paula (2004).

As we know, the data-driven covariance modelling method using modified cholesky decomposition can reproduce a wide range of classical, stationary and non-stationary, covariance structure. It is demonstrated that the method succeeds when the data has strong nonstationary covariance structure (see cattle data analysis in Pourahmadi 1999 and Pan and MacKenzie

2003). As a try, we apply the covariance modelling method to the diabetic patient data under the constrained model and compare it with those stationary covariance structures.

Pourahmadi’s model is used to estimate generalised autoregressive parameters, ϕ_{ijk} , $k < j$, $j = 2, \dots, 8$, and innovation variances, σ_{ij} , $j = 1 \dots, 8$. Noticing from below the main diagonal in Table 1 that the sample generalised autoregressive parameters vary among the indices j and k , we introduce the following model. For $j = 2, \dots, 8$ and $k = 1, \dots, 7$,

$$\begin{aligned} \phi_{jk} &= \gamma_0 + \gamma_1 j + \dots + \gamma_{q_1} j^{q_1} + \gamma_1^* k + \dots + \gamma_{q_2}^* k^{q_2} \\ &+ \gamma_1^{**} (j \times k) + \dots + \gamma_{q_3}^{**} (j \times k)^{q_3}, \end{aligned}$$

where $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_{q_1}, \gamma_1^*, \dots, \gamma_{q_2}^*, \gamma_1^{**}, \dots, \gamma_{q_3}^{**})'$ with $q = q_1 + q_2 + q_3$. In the rest of this section, Pourahmadi’s model is referred as *CM1* and our model is referred as *CM2*. For both models, we fit them twice using different degrees of polynomial. *CM1_a* fits the cubic functions of the lag to both the generalised autoregressive parameters and the innovation variances and *CM2_a* is the model with $q_1 = 2, q_2 = 2, q_3 = 2$ for the generalised autoregressive parameters and $d = 3$ for the innovation parameters. These models are roughly chosen by the principle of parsimony. It may be more standard to base our model choice on the *BIC* criterion, defined as $BIC = -(2/n)\widehat{l}_{max} + (q + d + 2) \log n/n$. We also give the *AIC* which is defined as $AIC = -(2/n)\widehat{l}_{max} + 2(q + d + 2)/n$ for comparison. Here n is the sample size and \widehat{l}_{max} is the maximized likelihood for a covariance model with the number of $(q + d + 2)$ parameters. The smaller the *AIC* or *BIC* value, the better the model. *CM1_b* gets the smallest value when $q = 5$ and $d = 7$ while *CM2_b* gets its smallest value when $q_1 = 5, q_2 = 4, q_3 = 1$ and $d = 3$. The sample generalised autoregressive parameters (below the main diagonal in Table 1 and diagram (a) in Figure 1) indicate that it is not good enough to estimate these parameters by using *CM1* which only fits the polynomial function of lag, i.e., $j - k$, even though the higher degree ($q = 5$) is used. It may be more reasonable to consider both the indices j and k in the regression models. This is done by *CM2*. The fitted regressograms using *CM2_b* are given in Figure 2. From the results in Table 2, *CM2* is a little better than *CM1* in the sense of *AIC* and *BIC*.

Table 2 also gives the comparison between menu selection method and covariance modelling. For compound symmetry, AR(1) and ARMA(1,1), Fisher-scoring algorithm with constrained optimization are applied to these models. It can be concluded from the table that the covariance modelling method still works well for the strong correlated data with stationary covariance structure. Furthermore, this conclusion is supported by comparison of sample and fitted correlations, generalised autoregressive parameters and innovation variances from Table 1 and Table 3.

TABLE 1. Diabetic patient data. Sample variances(along the main diagonal), correlations(above the main diagonal), generalised autoregressive parameters(below the main diagonal) and innovation variances(last row)

t	1	2	3	4	5	6	7	8
1	10	0.976	0.970	0.980	0.939	0.935	0.897	0.888
2	0.970	10	0.967	0.957	0.908	0.909	0.829	0.821
3	0.530	0.406	9	0.967	0.931	0.933	0.869	0.860
4	0.667	-0.198	0.533	10	0.958	0.954	0.893	0.889
5	0.185	-0.183	-0.034	1.069	12	0.967	0.935	0.929
6	-0.033	0.004	0.091	0.305	0.627	11	0.907	0.902
7	1.795	-1.229	0.360	-0.912	1.027	0.113	18	0.983
8	-0.050	-0.034	-0.109	0.219	0.064	0.000	0.936	19
	10.1	0.5	0.5	0.3	0.9	0.6	1.5	0.6

TABLE 2. Data-driven, regression-based, covariance modelling for Σ_i compared with several menu-selection methods

Structure of Σ_i	No. of parameters	\hat{l}_{max}	AIC	BIC
Compound symmetry	2	-310.26	29.74	29.84
AR(1)	2	-269.46	25.85	25.95
ARMA(1,1)	3	-267.77	25.79	25.94
$CM1_a$	7	-262.28	25.74	26.14
$CM2_a$	11	-258.26	25.64	26.19
$CM1_b$	14	-250.11	25.15	25.85
$CM2_b$	15	-247.11	24.96	25.71

TABLE 3. Diabetic patient data. Fitted innovation variances(along the main diagonal), correlations(above the main diagonal), and generalised autoregressive parameters(below the main diagonal) using $CM2_b$

t	1	2	3	4	5	6	7	8
1	9.7	0.958	0.977	0.972	0.942	0.886	0.902	0.896
2	0.969	0.8	0.959	0.941	0.892	0.819	0.833	0.810
3	0.663	0.265	0.3	0.973	0.940	0.879	0.890	0.874
4	0.471	-0.023	0.556	0.4	0.957	0.904	0.913	0.901
5	0.320	-0.269	0.214	0.756	0.8	0.919	0.922	0.912
6	0.308	-0.378	0.009	0.456	0.600	1.6	0.913	0.899
7	0.464	-0.317	-0.025	0.326	0.374	0.406	1.9	0.974
8	0.531	-0.346	-0.151	0.105	0.057	0.006	0.852	0.8

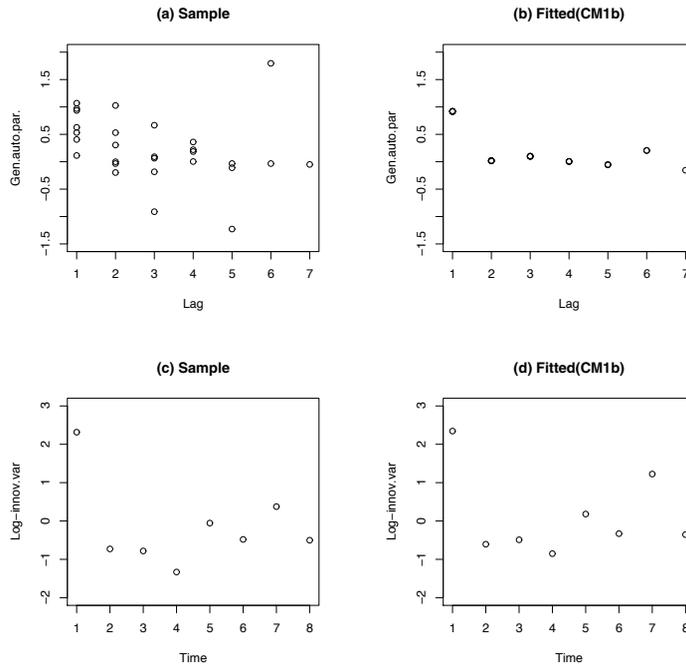


FIGURE 1. Sample and fitted regressograms for the diabetic patient data. (a) Sample generalised autoregressive parameters, (b) fitted generalised autoregressive parameters using $CM1_b$, (c) sample log-innovation variances, (d) fitted log-innovation variances using $CM1_b$.

Acknowledgments: Special Thanks to the Science Foundation Ireland (SFI) who provided the funding for this work.

References

Pan J., 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.

Xu J., Wang J. (2008). Maximum likelihood estimation of linear models for longitudinal data with inequality constraints. *Communications in Statistics-Theory and Methods*. **37**, 931-946.

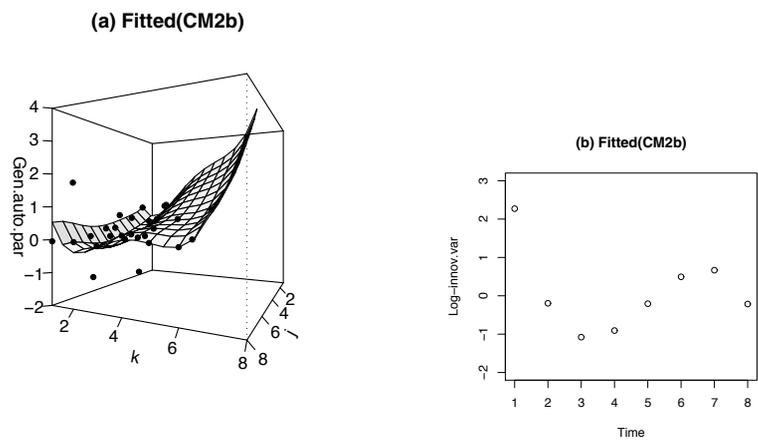


FIGURE 2. Fitted regressograms for the diabetic patient data. (a) Fitted generalised autoregressive parameters using $CM2_b$, (b) fitted log-innovation variances using $CM2_b$.

Author Index

- Adell, N., 295
- Amaral-Turkman, M.A., 49
- Aoki, Reiko, 322
- Babu, G. Jogesh, 237
- Bailey, T. C., 125
- Bar, Haim Y., 53
- Barber, Jarrett J., 174
- Blas Achic, Betsabé G., 61
- Bolfarine, Heleno, 61
- Booth, James G., 118
- Bowman, A.W., 69
- Brown, Jennifer, 168
- Caballero-Águila, R., 75
- Caffo, B., 3
- Caja, G., 295
- Camarda, Carlo G., 81
- Carné, S., 295
- Chakrabarty, Dalia, 89
- Ciera, James M., 96
- Colombi, Roberto, 102
- Costain, Deborah A., 110
- Cunningham, Caitlin M., 118
- Dasu, Tamraparni, 209
- de Castro, Mário, 138
- Dean, N., 69
- Declerck, Dominique, 39
- Djuraš, Gordana, 255
- Doerge, R.W., 13
- Dunson, David B., 96
- Durbán, María, 202
- Dvorzak, Michaela, 261
- Economou, T., 125
- Eilers, Paul H.C., 81, 130, 306, 330, 351
- Fabio, Lizandra C., 138
- Ferguson, C., 69
- French, Pim J., 306
- Gampe, Jutta, 81, 351
- Giordano, Sabrina, 102
- Greven, Sonja, 142
- Gross, J., 69
- Guerrero, Víctor M., 345
- Haines, Linda M., 150
- Hanlon, Bret M., 154
- Hansen, Bettina E. , 187
- Hedlin, Haley, 3
- Heller, Gillian Z., 160
- Hermoso-Carazo, A., 75
- Hodge, Miriam, 168
- Hong, Seung-Man, 281
- Huzurbazar, S., 174
- Jajo, N.K., 182
- Janssen, Harry L.A., 187
- Joel, Suresh, 3
- Jones, Chris, 267
- Kapelan, Z., 125

- Kaplan, Ray M., 154
Kauerman, Göran, 23
Kim, Kyunga, 13
Kneib, Thomas, 142
Kohn, Robert, 314
Komárek, Arnošt, 187
Kosmidis, Ioannis, 193
Krishnan, Shankar, 209
Lachos, Hugo, 61
Leask, Kerry L., 150
Lebarbier, Émilie, 243
Lee, Dae-Jin, 202
Leiva, Víctor, 365
Lesaffre, Emmanuel, 39, 183, 269
Letón, Emilio, 247
Lin, Dongyu, 209
Linares-Pérez, J., 75
MacKenzie, Gilbert, 217, 369
Marra, Giampiero, 223, 300
Martín, Nirian, 231
Martínez-Gómez, Elizabeth, 237
Mary-Huard, Tristan, 243
Matawie, K.M., 182
Matthews, Fiona E., 359
Meulman, Jacqueline J., 306
Mofstovsky, Stewart, 3
Molanes-López, Elisa-María, 247
Molas, M., 273
Neubauer, Gerhard, 255, 261
Newell, John, 351
Nielsen, Martin K., 154
Noufaily, Angela, 267
Oskrochi, G., 273
Pardo, Leandro, 231
Park, Hyo-Il, 281
Paula, Gilberto A., 138, 322, 365
Pekar, Jim, 3
Peña, Daniel, 345
Petersen, Stig L., 154
Pramanik, Chancal, 289
Puig, P., 295
Radice, Rosalba, 300
Reale, Marco, 168
Rigby, Robert A., 160
Rippe, Ralph C.A., 306
Robin, Stéphane, 243
Rojas-Olivares, A., 295
Rosen, Ori, 314
Russo, Cibele M., 322
Salama, A.A.K., 295
Scarpa, Bruno, 96
Schifano, Elizabeth D., 53
Schnabel, Sabine K., 330
Sellers, Kimberly F., 337
Shamley, D., 273
Shmueli, Galit, 337

Silva, Eliud, 345
Simpkin, Andrew, 351
Spear-Basset, Susan, 3
Stasinopoulos, D. Mikis, 160
Turkman, K.F., 49
van den Hout, Ardo, 359
van Ogtrop, Floris F., 160
Vidyshankar, Anand N, 154
Villegas, Cristian, 365
Vitolo, R., 125
Wagner, Helga, 261
Waterhouse, E., 125
Wood, Sally, 314
Wood, Simon N., 223
Xu, Jing, 217, 369