Proceedings of the
31st International Workshop on Statistical Modelling

Volume II

July 4–8, 2016
Rennes, France

Jean-François Dupuy, Julie Josse
(editors)

Proceedings of the 31st International Workshop on Statistical Modelling
volume II,
Rennes, July 4 – 8, 2016,
Jean-François Dupuy, Julie Josse (editors),
Rennes, 2016.

Editors:

Jean-François Dupuy, Jean-Francois.Dupuy@insa-rennes.fr
Department of Mathematical Engineering
Institut National des Sciences Appliquées Rennes
20 avenue des Buttes de Coësmes
35708 Rennes cedex 7, France

Julie Josse, Julie.Josse@agrocampus-ouest.fr
Department of Applied Mathematics
Agrocampus Ouest
65 rue de Saint-Brieuc
CS 84215
35042 Rennes Cedex, France

Scientific Programme Committee

- Avner Bar Hen
  Paris Descartes University, France
- Francesco Bartolucci
  Perugia University, Italy
- Dankmar Böhning
  Southampton University, UK
- Jean-François Dupuy
  INSA Rennes, France
- Jochen Einbeck
  Durham University, UK
- Marco Grzegorczyk
  Groningen University, Netherlands
- Ardo van den Hout
  University College London, UK
- Julie Josse
  Agrocampus Ouest, France
- Benoit Liquet
  University of Queensland, Australia
- Gerhard Tutz
  Munich University, Germany
- Helga Wagner
  Johannes Kepler University Linz, Austria

Local Organizing Committee

- Pierrette Chagneau (INSA Rennes)
- Jean-François Dupuy (INSA Rennes, Chair)
- Martine Fixot (INSA Rennes)
- Sabrina Jonin (INSA Rennes)
- Julie Josse (Agrocampus Ouest)
- Nathalie Krell (Rennes 1 University)
- James Ledoux (INSA Rennes)
- Audrey Poterie (INSA Rennes)
- Laurent Rouvière (Rennes 2 University)
- Myriam Vimond (ENSAI)

Contents

Part III – Posters

Part III - Posters

Penalized Gaussian Copula Graphical Models for Detecting Epistatic Selection

Pariya Behrouzi[1], Ernst C. Wit[1]

[1] Johann Bernoulli Institute, University of Groningen, Netherlands

E-mail for correspondence: P.Behrouzi@rug.nl

Abstract: The detection of high-dimensional epistatic selection is an important goal in population genetics. The reconstruction of the signatures of epistatic selections during inbreeding is challenging as multiple testing approaches are under-power. Here we develop an efficient method for reconstructing an underlying network of genomic signatures of high-dimensional epistatic selection from multi-locus genotype data. The network reveals "aberrant" marker-marker associations that are due to epistatic selection. The estimation procedure relies on penalized Gaussian copula graphical models.

Keywords: Epistasis; Gaussian copula; High-dimensional inference.

1 Introduction

Recombinant Inbred Lines (RILs) study design is a popular tool for studying the genetic and environmental basis of complex traits. RILs typically are derived by crossing two inbred parents followed by repeated generations to produce an offspring whose genome is a mosaic of its parental lines. Assuming genotype of parent 1 is labeled AA and that of parent 2 is labeled BB, the routine way of coding the genotype data is to use $\{0, 1, 2\}$ to represent $\{AA, AB, BB\}$, respectively. The construction of RILs is not always straightforward: low fertility, or even complete lethality, of lines during inbreeding is common. These genomic signatures are indicative of epistatic selection having acted on entire networks of interacting loci during inbreeding.

The reconstruction of multi-loci interaction networks from RIL genotyping data require methods for detecting genomic signatures of high-dimensional epistatic selection. One commonly used approach is hypothesis testing. The

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

drawback with such an approach is that hypothesis testing at the genome-scale is typically underpowered. Furthermore, theory shows that pair-wise tests are not, statistically speaking, consistent when two interacting loci are conditionally dependent on other loci in the network, and may therefore lead to an incorrect signatures.
In order to overcome some of these issues, we argue that the detection of epistatic selection in RIL genomes can be achieved by inferring a high-dimensional graph based of conditional dependency relationships among loci. Technically, this requires estimating a sparse precision matrix from a large number of discrete ordinal marker genotypes, where the number of markers p can far exceed the number of individuals n.

2 Methods

Multivariate Gaussian copula graphical model.
Let $Y_j^{(i)}$ $j = 1, \dots, p$; $i = 1, \dots, n$ denotes the genotype of ith individual for jth marker. The observations $Y_j^{(i)}$ arise from $\{1, \dots, K_j\}, K_j \geq 2$ discrete ordinal values. In the genetic set-up, K_j is the number of possible genotypes at locus j. A way to define the conditional independence relationships among the genetic markers is to assume an underlying continua for the $Y_j^{(1)}, \dots, Y_j^{(n)}$, which can not observed directly and which manifest itself through an ordinal scale of fixed length. In our modeling framework, $Y_j^{(i)}$ and $Z_j^{(i)}$ define the observed genotype and latent state, respectively. Each latent variable corresponds to the observed variables, expressed by a set of cut-points $(-\infty, C_1^{(j)}], (C_1^{(j)}, C_2^{(j)}] \dots, (C_k^{(j)}, \infty)$, which is obtained by partitioning the range of Z_j into K_j disjoint intervals. Generally, $y_j^{(i)}$ can be written as follow $\mathbf{y_j^{(i)}} = \sum_{\mathbf{k=1}}^{\mathbf{K_j}} \mathbf{k} \times \mathbf{1}_{\{\mathbf{c_{k-1}^{(j)}} < \mathbf{z_j^{(i)}} \leq \mathbf{c_k^{(j)}}\}}$.
Suppose the continuous latent variable Z has a p-variate normal distribution with $N_p(0, \boldsymbol{\Theta}^{\mathbf{-1}})$. The Gaussian copula modeling can be expressed as $Y_j = F_j^{-1}(\Phi(Z_j))$ where F_j represents marginals.
Inference of Gaussian copula graphical model.
Epistasis networks are known to be sparsely connected, so we impose sparsity on the elements of the precision matrix using an ℓ_1-norm penalty. We introduce the penalized EM algorithm for the estimation procedure. In the E-step the conditional expectation of the joint penalized log-likelihood given the data and $\widehat{\boldsymbol{\Theta}}^{(m)}$ can be determined as follow

$$Q_\lambda(\boldsymbol{\Theta} \mid \widehat{\boldsymbol{\Theta}}^{(m)}) = \frac{n}{2}[\log|\boldsymbol{\Theta}| - tr(\bar{R}\boldsymbol{\Theta}) - p\log(2\pi)] - \lambda||\Theta||_1 \qquad (1)$$

Here $\bar{R} = \frac{1}{n}\sum_{i=1}^{n} E(z^{(i)}z^{(i)t} \mid z^{(i)} \in \mathcal{D}(y^{(i)}), \widehat{\boldsymbol{\Theta}}^{(m)})$. The M-step involves the optimizing this conditional expectation. To obtain the $\bar{R}$, we use two simplifications

$$E(z_k^{(i)} z_l^{(i)t} \mid y^{(i)}, \widehat{\boldsymbol{\Theta}}) \approx \begin{cases} E(z_k^{(i)} \mid y^{(i)}, \widehat{\boldsymbol{\Theta}}) E(z_l^{(i)} \mid y^{(i)}, \widehat{\boldsymbol{\Theta}}) & \text{if } 1 \leq k \neq l \leq p \ ; \\ E(z_k^{(i)^2} \mid y^{(i)}, \widehat{\boldsymbol{\Theta}}) & \text{if } k = l. \end{cases}$$

Thus, the variance elements in the conditional expectation matrix can be calculated through the second moment of the conditional $z_j^{(i)} \mid y^{(i)}$, and the rest of the elements in this matrix can be approximated through the first moment of the truncated multivariate Gaussian distribution. To obtain the optimal model in terms of graph estimation we pick the penalty term that minimizes EBIC over $\lambda > 0$.

TABLE 1. Comparison the two methods over 50 independent run, where $p = 150, n = 50$. The best model in each column is boldfaced.

	Random			Scale-free		
	k=2	k=5	k=10	k=2	k=5	k=10
Approx						
F_1	**0.10**	**0.23**	**0.25**	**0.05**	**0.12**	**0.14**
SEN	0.30	0.22	0.30	0.20	0.18	0.15
SPE	0.90	0.98	0.99	0.90	0.97	0.99
NPNtau						
F_1	0.0	0.07	0.07	0.002	0.03	0.04
SEN	0.00	0.77	0.80	0.04	0.67	0.56
SPE	1.00	0.58	0.58	0.97	0.55	0.64

3 Data analysis

Simulations. We set up simulations to generate sparse matrices $\boldsymbol{\Theta}$ under commonly encountered genetic network structures: random and scale-free networks. We compare the performance of our proposed method with the nonparanormal skeptic Kendall's tau Liu *et al.* (2012) approach. We compare the two models for $n = 50$, $p = 150$, and different scenarios of $k \in \{2, 5, 10\}$. The results of the comparisons are provided in Table 1. We note that high values of the F_1-score, sensitivity (SEN) and specificity (SPE) indicate good performance. These results suggest that, though recovering sparse network structure from discrete data is a challenging task, the proposed approach performs well.

Epistatic selection in *Arabidopsis thaliana*. We apply our proposed method to detect epistatic selection in a RIL cross derived from A.*thaliana* genotype, where 367 individuals were genotyped for 90 genetic markers (M. Simon *et al.* (2008)). The A.*thaliana* genome has 5 chromosomes. Each chromosome contains 24, 14, 17, 16, 19 markers, respectively. Figure 1 shows that the bottom of chromosome 1 and the top of chromosome 5

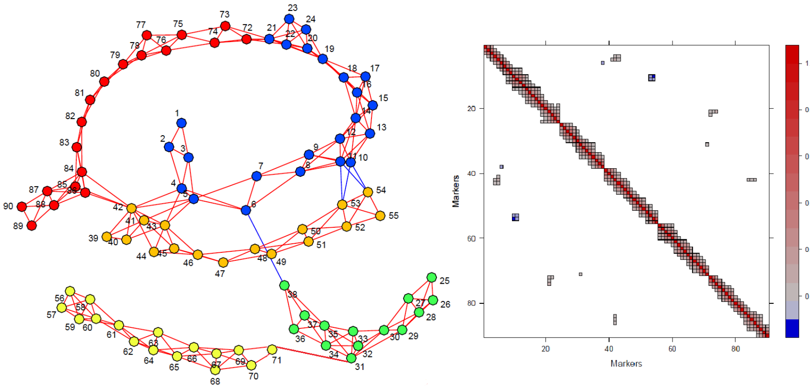


FIGURE 1. Inferred network for the genotype data in *A.thaliana*.

do not segregate independently from each other. Beside this, interactions between the top of chromosomes 1 and 3 involve pairs of loci that also do not segregate independently. This genotype has been studied extensively in Bikard *et al.* (2009). They reported that the first interaction we found causes arrested embryo development, resulting in seed abortion, whereas the latter interaction causes root growth impairment. In addition of these two regions, we have discovered few other trans-chromosomal interactions in the A.*thaliana* genome. These additional interactions may reveal other disorders in this crop that affect its viability.

4 Conclusion

The detection of high-dimensional epistatic selection is an important goal in population genetics. Our proposed method combines the Gaussian copula with Gaussian graphical models to explore the conditional dependencies among large numbers of genetic loci in the genome. Our simulations show that the proposed method outperforms the alternative method in terms of graph recovery. In the application of our method in the Arabidopsis, we discovered two regions that interact epistatically, in which cause arrested embryo development and root growth impairments.

References

D. Bikard *et al.* (2009). Divergent evolution of duplicate genes leads to genetic incompatibilities within A. thaliana. *Science*, 323 (5914), 623 – 626.

H. Liu *et al.* (2012). High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40 (4), 2293 – 2326.

M. Simon *et al.* (2008). QTL mapping in five new large RIL populations of Arabidopsis thaliana genotyped with consensus SNP markers. *Genetics*, 178, 2253 – 2264.

Spatio-temporal modelling of crime using low discrepancy sequences

Paul Brown[1], Chaitanya Joshi[1], Stephen Joe[1], Nigel McCarter[2]

[1] Department of Mathematics and Statistics, University of Waikato, Hamilton, New Zealand
[2] NZ Police Intelligence, Waikato, New Zealand

E-mail for correspondence: ptb2@students.waikato.ac.nz

Abstract: We perform spatio-temporal modelling of burglary data in order to predict areas of high criminal risk for local authorities. We wish to account for several spatio-temporal factors as latent processes to make the model as realistic as possible, thus creating a model with a large latent field with several hyperparameters. Analysis of the model is done using Integrated Nested Laplace Approximations (INLA) (Rue et al. 2009), a fast Bayesian inference methodology that provides more computationally efficient estimations than Markov Chain Monte Carlo (MCMC) methods.

Keywords: Bayesian inference; Crime modelling; Integrated nested Laplace approximations; Low discrepancy sequences; Spatio-temporal modelling

1 Introduction

Efficient use of police resources is vital for taking preventative measures against crime, as opposed to reactive measures. As such, intelligence-led policing, where data, analysis, and criminal theory are used to guide police allocation and decision-making, are becoming ever more popular and necessary (Ratcliffe, 2012). Given that crime and its associated factors occur within a geographical context that include both space and time (Fitterer et al., 2014) spatio-temporal modelling can lead to more accurate prediction of crime than modelling which does not take these factors into consideration. Spatio-temporal modelling under the Bayesian paradigm also allows more information to be included through prior knowledge from local authorities and officials.

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

2 Data and Methods

2.1 Burglary Data

The main dataset consists of residential, petty (under NZ$5000) burglaries from 2010 to 2015 in the Hamilton City region, New Zealand. All locations are geocoded using the New Zealand Transverse Mercator (NZTM) northings and eastings. The region is bounded by an 11 kilometre (km) × 13 km rectangle, and is partitioned up in 1 km × 1 km cells, giving 143 cells in total (see Figure 1).

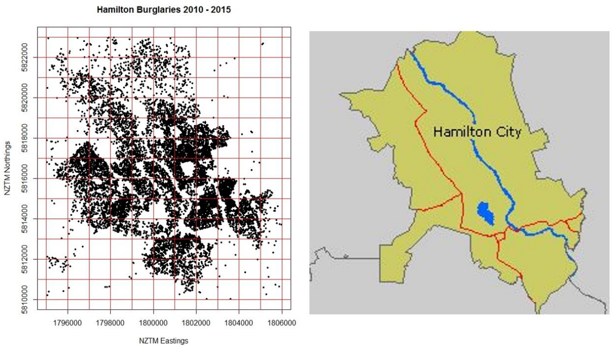


FIGURE 1. Hamilton burglaries from 2010 to 2015 and map of Hamilton, NZ.

There are several underlying factors involved with the spatial distribution of crime. Substantial research indicates that certain segments of the population and particular types of environments can generate high offender rates (Brown, 1982). The New Zealand Index of Deprivation (NZDEP) is a measure of socioeconomic status of a small geographical area, and includes a range of variables including income, employment and living spaces (see Atkinson et al. (2014) for full details). The social and physical environment is represented by off-licence liquor stores and graffiti respectively. Off-licence liquor stores, which may encourage a higher consumption of alcohol is also used as a measure of the perception of lawlessness and also of low anti-social behaviour. Incidence of graffiti has been used as a measure of lawlessness and environmental deterioration within an area.

2.2 INLA

INLA is a fast Bayesian inference methodology for latent Gaussian models that take the form

$$\eta_i = \beta_0 + \sum_{j=1}^{n_\beta} \beta_j z_{j,i} + \sum_{k=1}^{n_f} f^{(k)}(u_{k,i}) + \epsilon_i,$$

where the $\{\beta_k\}$'s are the linear effect on covariates $\mathbf{z}$, $\{f^{(\cdot)}\}$'s represent the unknown functions of the covariates $\mathbf{u}$, and ϵ_i's are the unstructured random errors. The latent Gaussian model is obtained by assigning all parameters in the latent field $\boldsymbol{\phi} = \{\beta_0, \{\beta_k\}, \{f^{(\cdot)}\}, \{\eta_i\}\}$ a Gaussian prior with mean $\mathbf{0}$ and precision matrix $Q(\boldsymbol{\theta})$, with hyperparameters $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_m\}$. Inference on $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are made via the use of numerical integration, Laplace approximations and grid sampling strategies. For a full account of the inference stage, see Rue et al. (2009).

2.3 Modelling

Let $y_{i,t}$ be the count of burglaries in cell i at time t. Assume that $y_{i,t} \sim \text{Poisson}(\lambda_{i,t})$, and $\log(y_{i,t}) = \eta_{i,t}$. We have a generalised additive model

$$\eta_{i,t} = \beta_0 + \beta_1 \text{NZDEP}_i + r_i + g_i + l_i + Y_t + \epsilon_{i,t},$$

where β_0 is the overall mean, β_1 is a fixed linear parameter for the covariate NZDEP. The term r_i is the spatial effect of each cell, and is modelled as a Gaussian Markov Random Field (GMRF) with unknown precision τ_r.This specification, also known as conditionally autoregressive (CAR) prior, was introduced by Besag et al. (1991) and is used extensively in disease mapping. The terms g_i and l_i are incidence of graffiti and number of liquor stores respectively and are both modelled similarly to r_i, with hyperparameters τ_g and τ_l. The term Y_t represents the yearly time effect and is modelled as a random walk of order 1 (RW1) and has a hyperparameter τ_Y. Hence our latent parameters are $\boldsymbol{\phi} = \{\{\eta_{i,t}\}, \beta_0, \beta_1, \{r_i\}, \{g_i\}, \{l_i\}, \{Y_t\}\}$ with hyperparameters $\boldsymbol{\theta} = \{\tau_r, \tau_g, \tau_l, \tau_Y\}$.

3 Preliminary Results

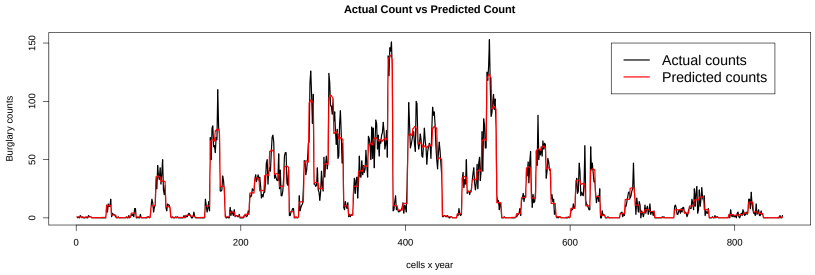


FIGURE 2. Actual count of burglaries per year vs. predicted count using INLA.

Figure 2 shows that the model predicts the actual counts well. Note that there are 143 cells, each with six years of data, giving the total of 858 cells for prediction. There is some variability between the years for each cell. There may be other temporal factors, such as seasonality, that may play an important role. Adding and developing the model, as well as model performance is currently being worked on.

4 Ongoing Work

There are many factors that may be involved in burglaries that we have not considered yet. Many of these factors may have some spatial or temporal effects, thus adding to the number of latent and hyper-parameters in the model. INLA could lose its computational efficiency for models with a large number of hyperparameters, therefore we need a methodology that can provide accurate and efficient inference on such a model. Recently, a paper by Joshi et al. (2016) has proposed using low discrepancy sequences instead of grids at the hyperparameter stage to increase computational gains and accuracy. In ongoing work, we are building a larger model with many hyperparameters using this approach.

References

Atkinson, J., Salmond, C., and Crampton, C. (2014). *NZDEP2013 Index of deprivation.* NZ Ministry of Health, http://www.health.govt.nz

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

Brown, M.A. (1982). Modelling the spatial distribution of suburban crime. *Economic Geography*, **58:3**, 247 – 261.

Fitterer, J., Nelson, T.A., and Nathoo, F. (2014). Predictive crime mapping. *Police Practice and Research: An International Journal*, **16:2**, 121 – 135.

Joshi, C., Brown, P.T., and Joe, S. Fast Bayesian inference using low discrepancy sequences. *Bayesian Analysis*, (Submitted).

Ratcliffe, J. (2012). *Intelligence-led policing.* Portland, OR: Willian Publishing.

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*, **71:2**, 319 – 392.

The memory of extrapolating P-splines

Alba Carballo[1], María Durbán[1], Dae-Jin Lee[2], Paul Eilers[3]

[1] Department of Statistics, Universidad Carlos III de Madrid, Spain
[2] Basque Center for Applied Mathematics, Bilbao, Spain
[3] Department of Biostatistics, Erasmus University Medical Center, Rotterdam, The Netherlands

E-mail for correspondence: albcarba@est-econ.uc3m.es

Abstract: There are many situations in which forecasting with smoothing models is needed, for example, hourly temperatures at a weather station or yearly number of deaths. Due to the kind of data and to the information available it may be important to know how much of the past information we are using to forecast. We introduce the concept of *memory of a P-spline* as a tool to provide that information and show some of its properties. We illustrate the concept with a data set of mortality of Spanish men.

Keywords: Forecasting; P-splines.

1 Forecasting with P-splines

Consider the case of a univariate Gaussian data, with ordered regressor $\boldsymbol{x}$ and response variable $\boldsymbol{y}$. The smooth model is of the form:

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}),$$

where $f(\cdot)$ is an unknown smooth function and $\boldsymbol{\epsilon}$ are independent and identically distributed errors with variance σ^2. Suppose, given n observations $\boldsymbol{y}$ of the response variable, that we want to predict n_p new values $\boldsymbol{y}_p$ at $\boldsymbol{x}_p$. Currie et al. (2004) proposed a method for fitting and forecasting simultaneously with smoothing models, it is based on the smoothing technique of penalized splines proposed in Eilers and Marx (1996), i.e., the basic idea is to use a B-splines basis as the regression basis and modify the likelihood function by adding a penalty term over adjacent regression coefficients to control the smoothness of the fit.
Currie et al. (2004) construct a B-spline basis $\boldsymbol{B}$ from a set of knots which range covers all values of $\boldsymbol{x}_+ = (\boldsymbol{x}', \boldsymbol{x}_p')'$. The B-spline basis has size $n_+ \times c$, with $n_+ = n + n_p$ and c the length of the vector of coefficients, $\boldsymbol{\theta}$.

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

To obtain simultaneously the fit and the forecast Currie et al. (2004) minimize the following function of $\boldsymbol{\theta}$:

$$S = (\boldsymbol{y}_+ - \boldsymbol{B\theta})'\boldsymbol{M}(\boldsymbol{y}_+ - \boldsymbol{B\theta}) + \lambda\boldsymbol{\theta}'\boldsymbol{D}'\boldsymbol{D\theta},$$

where $\boldsymbol{M}$ is a diagonal weight matrix of size $n_+ \times n_+$ with diagonal elements equal to 0 if the data is missing or forecasted and 1 if the data is observed, $\boldsymbol{y}_+ = (\boldsymbol{y}', \boldsymbol{y}_p')'$ is the vector of observations extended, it contains the observed response, $\boldsymbol{y}$, and arbitrary values, $\boldsymbol{y}_p$, λ is the smoothing parameter and $\boldsymbol{D}'\boldsymbol{D}$ is the penalty matrix. The penalized least square solution give the fit and the forecast:

$$\hat{\boldsymbol{y}}_+ = \boldsymbol{H}_+\boldsymbol{y}_+,$$

with $\boldsymbol{H}_+$ the hat matrix, $\boldsymbol{H}_+ = \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{MB} + \lambda\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{B}'\boldsymbol{M}$.

Notice that because the last columns of $\boldsymbol{H}_+$ are all zeros (as the corresponding diagonal elements of $\boldsymbol{M}$ are zero), $\boldsymbol{H}_+$ has the following form:

$$\boldsymbol{H}_+ = \begin{bmatrix} \boldsymbol{H} & \boldsymbol{O}_1 \\ \boldsymbol{H}_p & \boldsymbol{O}_2 \end{bmatrix}, \tag{1}$$

with $\boldsymbol{H}$ of size $n \times n$, $\boldsymbol{H}_p$ of size $n_p \times n$ and $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$ matrices of zeros of size $n \times n_p$ and $n_p \times n_p$, respectively. Therefore, $\hat{\boldsymbol{y}} = \boldsymbol{Hy}$ and $\hat{\boldsymbol{y}}_p = \boldsymbol{H}_p\boldsymbol{y}$.

2 Memory of a P-spline

Using the previous approach, we have that $\hat{\boldsymbol{y}}_p = \boldsymbol{H}_p\boldsymbol{y}$ are the predicted values, i.e., the values of the rows of $\boldsymbol{H}_p$ give the influence of each observed value on the predicted values. Therefore the key point to know how much past information we are using to forecast is to summarize each row of $\boldsymbol{H}_p$ in a meaningful way. To simplify the notation we define $\boldsymbol{G}_p = \text{abs}(\boldsymbol{H}_p)$.

We have noticed that all rows of $\boldsymbol{G}_p$ follow a similar pattern, i.e., if we consider each row as a function and we study their monotony we find that this is the same for all the rows. For instance, if the maximum of the last row is taken at the last column, this also happens in the rest of columns. Moreover, in every row of $\boldsymbol{G}_p$ there are elements significantly larger than the others. The rows of a particular matrix $\boldsymbol{G}_p$ are plotted in the right panel of Figure 1.

Based on these ideas we have developed the concept *memory of a P-spline*, to know the overall weight of each observation on the prediction we have added the columns of $\boldsymbol{G}_p$ and consider these values (after dividing by their sum) as a vector of weights, $\boldsymbol{w}$. Considering the domain, T, as the number of steps backward, we define the memory of the P-spline as the 99^{th} percentile, t_0. Thus, t_0 is the number of steps backward we are taking information.

Notice that consider the memory as the 99^{th} percentile is just one possibility to summarize the vector of weights. Summary statistics that treat the weights as if they are a discrete distribution (mean, quantiles, expectiles) are other choices.

We performed a simulation study (which we do not include due to the lack of space) and we concluded:

- The memory does not depend on the prediction horizon.

- Depending on the variability of the observed data the memory is smaller or larger, i.e., the past has more or less influence on the predicted values. The smoother the trend is, the greater the influence of the past on the predicted values is. On the other hand, the rougher the trend is, the smaller the influence of the past on the predicted value is.
- The memory, like the effective dimension, only depends on the smoothing parameter and not on the size of the B-spline basis (provided that the basis is sufficiently large). Models with equal effective dimensions have similar memories.

3 Illustration

To illustrate the concept of *memory of a P-spline* we use a data on the log mortality rates of Spanish men aged 29 between 1960 and 2009. The data set contains 50 observations, i.e., the size of the hat matrix that give us the fit is 50 × 50. If we forecast up to 2019, i.e., we compute 10 new observations, the hat matrix $\boldsymbol{H}_p$ has size 10 × 50, the absolute value of the rows of $\boldsymbol{H}_p$ is plotted in the right panel of the Figure 1.

TABLE 1. Normalized weights, $\boldsymbol{w}_t$, for the number of steps backward from the last observed year.

t	$\boldsymbol{w}_t$	t	$\boldsymbol{w}_t$	t	$\boldsymbol{w}_t$
1	0.5034	7	0.0171	13	0.0005
2	0.0747	8	0.0016	14	0.0002
3	0.1039	9	0.0067	15	0.0004
4	0.1314	10	0.0066	16	0.0003
5	0.0961	11	0.0043	17	0.0002
6	0.0506	12	0.0020	18	0.0001

To calculate the memory of the P-spline, we consider the vector of weights, $\boldsymbol{w}$, containing the standardized sum of the columns of $\boldsymbol{G}_p$, its values are shown in Table 1 (the values of $\boldsymbol{w}_t$ for $t = 19, ..., 50$ are not shown in the table since they are approximately 0). In this case the memory of the P-spline, the 99^{th} percentile, is $t_0 = 9$, i.e., what has happened 9 years backward, before 2001, does not influence on the future. Figure 1 shows the fit and the forecast of the log mortality rates until 2019, the data that are between the red and the black lines correspond to the data that influence the prediction, the data associated to the years 2001-2009.

Acknowledgments: The work of the authors has been funded by the Ministry of Economy and Competitiveness grant MTM2014-52184-P. The research of Dae-Jin Lee was also supported by the Basque Government through the BERC 2014-2017 program and by the Spanish Ministry of Economy and Competitiveness MINECO: BCAM Severo Ochoa excellence accreditation SEV-2013-0323.

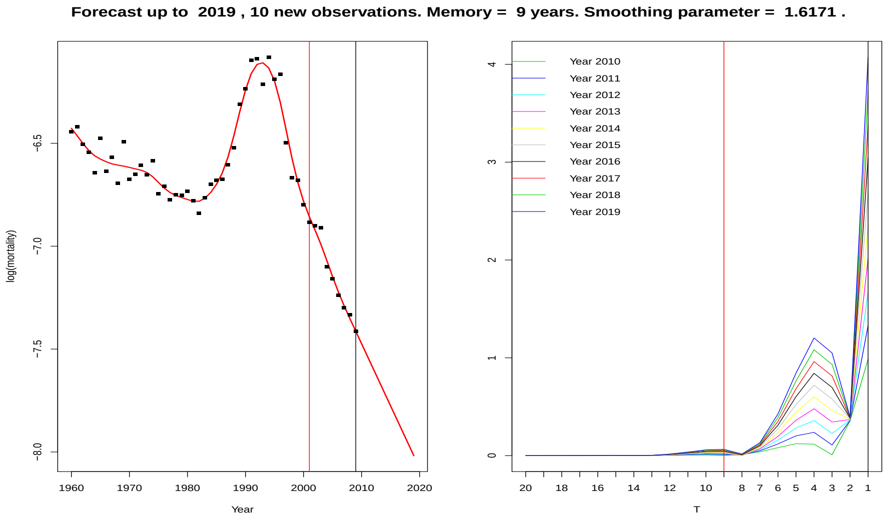


FIGURE 1. Left panel: fit and forecast of the log mortality rates until 2019, the data that there are between the red and the black lines correspond to the data that influence the prediction. Right panel: rows of $\boldsymbol{G}_p$, the red line corresponds to the number of backward steps we are taking information, 9.

References

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

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

Modelling censored and uncensored data: comparing treatments with GOF techniques

Claudia Castro-Kuriss[1], Víctor Leiva[2]

[1] Instituto Tecnológico de Buenos Aires, Argentina
[2] Universidad Adolfo Ibáñez, Chile

E-mail for correspondence: ccastro@itba.edu.ar

Abstract: Goodness-of-fit (GOF) techniques have been widely studied to assess the fit of distributions belonging to the location-scale (LS) family. However, several truncated and skewed distributions used in biostatistics belong to the non-location-scale family (NLS). In addition, because of time or money decisions, some epidemiological studies end before all the patients enrolled in the trial die or experience the event, producing censored data. Also, the comparison of two treatments is a well-known problem in medicine. One could be interested in comparing a new drug with the usual one or with a placebo. We analyze data from a clinical trial where patients were assigned to drug or placebo in a surgery intervention with uncensored observations. It is usually assumed that the underlying distributions in both populations are the same with different location parameters. We show that this is not always true, adjusting a LS distribution to the placebo group and a gamma mixture distribution to the other. We consider also real censored survival times finding that a NLS distribution best describe them. Our research provides appropriate tools and new perspectives in model selection using new GOF tests and graphical techniques. We illustrate the provided GOF results to the real-world data sets with probability plots that indicate a very good specification of the postulated hypothetical distribution under H_0.

Keywords: Censored Data;GOF tests;Location and Non-Location-Scale Family.

1 Introduction

Goodness-of-fit (GOF) tests have been developed for establishing the fitting of a distribution to a data set. In particular, in reliability and survival analysis, parametric life distributions are commonly used. GOF tests establish if the null hypothesis H_0 cannot be rejected based on empirical evidence. We have two possible options: (1) the distribution under H_0 is completely specified and (2) some

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

or all the parameters of the distribution are unknown. The most common case is the second one, where it is necessary to find a distribution in the family proposed under H_0 by means of proper parameter estimates in which case we suggest to use the maximum likelihood (ML) method. We focus on GOF statistics that measure the distance between the empirical cumulative distribution function (ECDF) and CDF established under H_0: Anderson-Darling (AD) and Kolmogorov-Smirnov (KS). Also, we consider Michael (MI) statistic based on a modification of the KS statistic using an arcsin transformation; see Michael (1983). The KS and MI statistics can be related to graphical plots, which show how well the specified theoretical distribution fits the data. Such graphs are the probability-probability (PP) and stabilized probability (SP) plots. GOF tests need to be adapted to censored samples. Several GOF tests for the LS family and different censored schemes are known. Some new available GOF tests for the NLS family can be considered as derivations from the proposal of Chen and Balakrishnan (1995), which were extended to type II censored samples and unknown parameters; see Castro-Kuriss et al. (2014). In this work, we propose to select models based on hypothesis testing to compare two different treatments. The model selection can be performed among those distributions where parameters with censored data can be estimated, regardless the family where the model is contained.

2 GOF tests for censored and uncensored data.

Consider the hypotheses: H_0: "the data are generated from a distribution with CDF $F(\cdot)$" versus H_1: "the data are not generated from this distribution". The hypothesized distribution with CDF $F(\cdot)$ can depend on a parameter vector θ containing location (μ), scale (β), shape (α) parameters or any other parameter not necessarily of location and scale. If the hypotheses of interest H_0 considers $F(t) = F([t-\mu]/\beta)$ with unknown parameters, we elaborate Algorithm 1 (A1) to perform the test, estimating properly the unknown parameters. Chen & Balakrishnan (1995) proposed an approximate GOF test that can be applied to NLS distributions. This method first transforms the data to normality and then applies A1 using the CDF of the normal model that allows us to compute the critical values of the corresponding test statistics, independently of the parameter estimators, if consistent estimators are available and the sample size exceeds 20. When the NLS family is considered under H_0 with unknown parameters, we propose Algorithm 2 (A2). GOF tests for NLS distributions with censored data can be obtained adapting the GOF statistics. Graphical plots with acceptance bands, like PP and SP plots, can also be derived to test NLS distributions under H_0 considering type II right censored samples.

3 Applications to real world data sets

3.1 Example 1: Comparison of two treatments.

Certain clinical trials are aimed at shortening the time-to-discharge. In a double-blind placebo controlled drug study, Shuster et al.(2008) reported times (in hours) of 23 patients on drug and of 25 patients on placebo. No censoring occurred on this trial. The hypothesis was that a 4-day ambulatory femoral nerve block decreases the lengthofstay after a total knee arthroplasty compared with the usual treatment. The placebo data show a distribution skewed to the right. Hence, we consider 12 possible distributions including the generalized Birnbaum-Saunders (GBS) with different kernels. The ML estimates and the corresponding observed statistics for the selected Weibull (WE) model are omitted, but $0.4 <$ p-value< 0.5 for the less powerful test. We notice that the drug data seem to be generated by two different populations. Only two models fit well the data: mixture normal and mixture gamma, being better the last one. We reject the WE model with a p-value< 0.001. We omit here ML estimates of the parameters from the drug group, components of the mixture gamma model and the observed statistics, but $0.25 <$p-value< 0.4 for the less powerful test. By means of the selected distributions, we estimate that 43.4% of the patients that received the conventional treatment and 4.4% of the patients that received the new drug stay in the hospital more than 3 days (the usual estimated time): the drug is excellent reducing the length of stay. Figure 1 shows the histogram and estimated PDF of the indicated distributions for placebo and drug data in different scales. We omit here the PP and SP plots for each group.

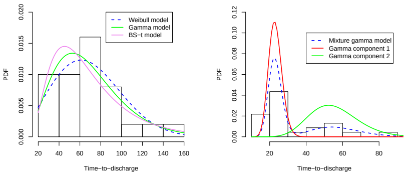


FIGURE 1. Histogram and estimated PDF of the indicated distribution for data in the placebo (left) and drug (right) groups.

3.2 Example 2: times to failure in an accelerated life test .

We analyze the times to failure in a temperature-accelerated life test for a device (Meeker and Escobar; 1998). The sample is singly censored to the right with 33 failures and 4 censored observations at 5000 hours. We evaluate the adequacy of 5 life distributions for which we can estimate their parameters in the case of a censored sample and then use A2. The GBS model with normal kernel is the distribution that provides the best fit to these data. The ML estimates of the GBS parameters and the obtained observed statistics are omitted here, but for the 3 of them we obtain $0.9 <$p-value< 0.95 indicating an excellent adequacy of the model. Figure 2 shows the PP and SP plots of the times to failure for device according to the selected model with 95% acceptance bands. As expected all the observations fall inside the bands with very good alignment.

4 Conclusions

We used the available tests for the NLS family and discussed the possibility of model selection in both LS and NLS families based on a hypothesis test approach.

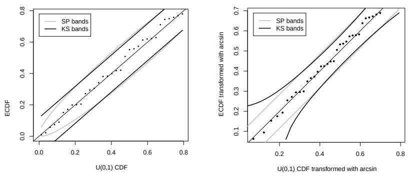


FIGURE 2. PP (left) and SP (right) plots with 95% acceptance bands for times to failure of a device using the BS model.

We proposed to use this approach to compare two or more treatments with uncensored or censored observations. GOF tests for mixture distributions under censoring schemes is under study by the authors.

References

Castro-Kuriss, C., Leiva, V. and Athayde, E. (2014). Graphical tools to assess goodness-of-fit in non-location-scale distributions. *Colombian Journal of Statistics (Special issue on "Current Topics in Statistical Graphics").* **37**, 341 – 365.

Chen, G. and Balakrishnan, N. (1995). A general purpose approximate goodness-of-fit test. *Journal of Quality Technology.* **27**, 154 – 161.

Meeker W. Q. and Escobar L. A. (1998). *Statistical Methods for Reliability Data.* New York: John Wiley & Sons.

Michael, J.R. (1983). The stabilized probability plot. *Biometrika.* **70**, 11 – 17.

Shuster, J., Theriaque, D. and Ilfeld B. (2008) Applying Hodges-Lehmann scale parameter estimates to hospital discharge times. *Clinical Trials.* **5**, 631 – 634.

Comparison of the SAM and a Bayesian method for differential gene expression analysis

Linda Chaba[1], John Odhiambo[1], Bernard Omolo[2]

[1] Strathmore Institute of Mathematical Sciences, Strathmore University, Kenya
[2] Division of Mathematics & Computer Science, University of South Carolina-Upstate, USA

E-mail for correspondence: lchaba@strathmore.edu

Abstract: Microarray technology has enabled the study of expression of thousands of genes simultaneously. One of the main objectives in microarray experiments is the identification of a panel of genes that are associated with a disease outcome or trait. Many statistical methods have been proposed for gene selection in the recent past but few systematic comparisons among these methods exist. A review and comparison of the statistical methods may provide biomedical researchers a useful guide for choosing the right method for a given microarray data in differential gene expression analysis.

This study reviewed a Bayesian method for the false discovery rate (FDR) control, based on the direct posterior probability approach, and the significance analysis of microarrays (SAM) method, and compared their performance when applied to two publicly available datasets from melanoma studies. The two approaches were compared in terms of the power to detect differential gene expression, the predictive ability of the genelists for a continuous outcome (G_2 checkpoint function), and the prognostic properties of the genelists for distant metastasis-free survival. Enrichment analysis was also performed to determine the biological usefulness of the genelists. The list generated by the SAM method contained fewer genes but performed better in terms of prediction and prognosis. The Bayesian approach was more powerful in detecting differential gene expression and contained more important genes in melanoma biology than the SAM method.

The SAM method is the most commonly used method in feature selection in microarray studies but loses power when the number of arrays (samples) are small. The Bayesian approach would be more suitable under these circumstances. However, we would recommend the SAM method to melanoma researchers, due to the predictive and prognostic properties of the genelist it generated.

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

Keywords: False discovery rate; Gene expression; Microarray

1 Introduction

Microarray technology has revolutionized genomic studies by enabling the study of differential expression of thousands of genes simultaneously. In the recent past, statistical methods have been developed to find differentially expressed genes. Tusher et al.(2001) proposed the SAM method, which identified genes with statistically significant changes in expression by assimilating a set of gene-specific t-tests. Smyth (2005) developed a method that fits a linear model to the expression data for each gene and uses Empirical Bayes and other shrinkage methods to borrow information across genes. Bayesian statistical methods have also been developed for differential gene expression with a view to, *inter italia*, finding significant genes or gene signatures in large oncological microarray studies. Among those who studied Bayesian approaches and their applications to microarray are Scharpf et al. (2009) and Lee et al. (2003).

Despite all the proposed methods mentioned above, there is no unanimous agreement on any particular gene selection method as the standard. However, some methods, like SAM, are more commonly used than others. A review and comparison of the statistical methods may provide bioinformaticians and other biomedical researchers a useful guide for choosing the right method for the right data in differential gene expression analysis. Furthermore, even though work has been done on the development of methods for the differential analysis of gene expression data measured in two conditions, open research questions still exist regarding the analysis of gene expression data in which the training signal is a continuous variable.

This paper reports a review of a Bayesian method for controlling the FDR and the SAM method and their comparison in identifying genes that are associated with a continuous outcome from the systems biology of melanoma, using a larger number of melanoma cell-lines than reported in Kaufmann et al.(2008). While the comparison of the SAM and the empirical Bayesian approaches to differential gene expression analysis has been done, no study to our knowledge has assessed the biological and clinical significance of genes identified by the SAM and the Bayesian method based on the direct posterior probability approaches. Our study has attempted to fill this gap in the literature. The comparison is based on the size and the statistical assessment of the predictive and the prognostic properties of the genelists produced by the two methods.

Material and Methods

Data

In our study, the gene expression data (raw intensities) consisted of 54 cell-lines (35 melanoma tumors and 19 normal human melanocytes), each with 41,093 probes. Only the melanoma tumors were analyzed. After filtration and normalization, 23,360 genes were available for analysis. The data matrix consisted of log2 ratios of ($G = 23,360$) genes on ($n = 35$) samples. An independent data set, consisting of gene expression data from 6307 genes on 58 primary melanomas with survival outcome, was also obtained for assessing prognosis. This data set has been reported in Winnepenninckx et al. (2006) and will hereafter be referred to as the Winnx dataset.

G_2 checkpoint function was selected to quantify the biological process in melanoma progression. The G_2 checkpoint is a position of control in the cell cycle that delays or arrests mitosis when DNA damage by radiation is detected. The G_2 checkpoint prevents cells with damaged DNA cell from entering mitosis, thereby providing the opportunity for repair and stopping the proliferation of damaged cells. Pathology experiments were conducted at Kaufmann's lab (UNC - Pathology and Lab Medicine) to assess the G_2 checkpoint function. For this study, the G_2 checkpoint function in melanoma cell-lines was scored as a ratio of mitotic cells in 1.5 Gy ironizing radiation (IR)-treated cultures in comparison to their sham-treated control (i.e. IR to sham ratio) (Omolo et al.(2013)). Melanoma cell-lines with G_2 scores greater than 0.30 were considered as checkpoint *defective*; otherwise they were *effective*.

Statistical Analysis

We applied SAM and Bayesian approach to find genes that are associated with G_2 checkpoint function. SAM employs the FDR control for the multiple testing problem and estimates the FDR through the permutation of values of the response variable and the gene-specific score while Bayesia approach permits control of the FDR using the direct probability approach. The two gene list generated by the two methods are hereafter be called SAMlist and Bayeslist respectively. In order to get additional insight into the performance of the two approaches, the two genelists were intersected to get the overlapping genes, hereafter known as SAMBayeslist.

We assessed the predictive quality of each of the genelists by their mean squared error (MSE) of prediction of the G_2 checkpoint function. For this, linear models containing significant genes were formulated. Since $G >> n$, the least absolute shrinkage and selection operator (LASSO) algorithm Tibshirani (1996) was used to select genes to include in the models. LASSO builds a sequence of models containing upto n genes and index by F, the number of algorithm steps relative to the model containing n genes (full model). For each F, a cross-validation estimate is obtained using leave-one-out cross-validation (LOOCV) method. The final model selected corresponds to the F-value with minimal estimated mean squared error.

To determine the clinical significance of the genelists, we performed supervised principal component analysis to identify genes that are significantly associated with a clinical outcome in the Winnx dataset. The clinical outcome for this dataset was 4-year distant metastasis-free survival (DMFS) and the objective was to predict a patient's risk (low/high) for developing distant metastasis within 4 years of primary diagnosis.Samples (patients) with a prognostic index above the median are classified as high risk; otherwise, they are low risk. A log-rank test is performed to test if the two survival curves for the low- and the high-risk groups are significantly different, using the original DMFS values. The power of the log-rank test is assessed through 1000 random permutations of the survival and censoring times. A genelist would be prognostic for DMFS if the log-rank test is significant. We compared the performance of the genelists produced by the two methods in survival risk prediction for the 58 samples in the Winnx dataset.

Results and Discussion

We refer to Table 1 below for a summary of our main results. The Bayesian approach identified 895 significant 895 (Bayes genelist) compared to 153 by the SAM approach (SAM genelist) at an FDR of 0.167411. The intersection of the two genelists yielded 129 overlapping genes, hereby referred to as the SAMBayes list. The three genelists were subjected to unsupervised hierarchical clustering analysis in order to assess the separation of the 35 melanoma cell-lines (samples) into G_2-defective and G_2-effective groups. Hierarchical clustering analysis was performed using BRB-ArrayTools version 4.4.1, (Simon et al.(2007)). The defective and effective cell-lines were also found to be statistically different ($W = 276, P < 0.01$). The naïve estimate of the mean square error was found to be 0.31 for the SAM genelist, 0.23 for the Bayes genelist and 0.29 for the SAMBayes genelist. The three genelists were also used to build linear predictive models for the G_2 checkpoint function, via the LASSO with LOOCV. The R^2 for the SAM genelist was 0.61 with a predictive accuracy of 91%, while the R^2 for the Bayesian genelist was 0.38 with a predictive accuracy of 69%. The SAMBayes genelist yielded an R^2 of 0.43 with a predictive accuracy of 83%. Gene expression data for the three genelists were extracted from the Winnx dataset for performing survival risk prediction. The difference between the survival curves for the low- and high-risk groups was significant for the SAM genelist ($\chi^2 = 7.5, P = 0.0235$, and the SAMBayes genelist ($\chi^2 = 4.4, P = 0.0363$, but not significant for the Bayes genelist ($\chi^2 = 1.8, P = 0.175$). Clearly, the SAM genelist was more accurate in the prediction of the G_2 checkpoint function and more prognostic for DMFS than the Bayes genelist. Annotation and enrichment analysis was performed for the three genelists using the Database for Annotation and Integrated Discovery (DAVID) version 6.7 (Huang(2008)).

TABLE 1. Comparison of G_2 checkpoint function prediction by the SAM, Bayesian, and SAMBayes genelists. The number of genes associated with DMFS (Cox genes) are also included.

	SAM	**Bayes**	**SAMBayes**
Genelist	153	895	129
Genes in Model	29	34	15
R-squared	0.61	0.38	0.43
Accuracy	91%	69%	83%
Cox Genes	26	151	20

Conclusion

The SAM and a Bayesian method for differential gene expression were compared in terms of their power to detect differential gene expression, the predictive abil-

ity of the genelists for a continuous outcome, and the prognostic properties of the genelists for DMFS. While the Bayesian approach was more powerful in terms of the number of significant genes detected, the genelist generated by the SAM approach performed better in terms of prediction and prognosis. The Bayes genelist was also enriched with lysosomal genes and contained other genes that are associated with regulation of cell cycle progression and melanomagenesis.
Based on our analysis, the SAM approach would be preferred over the Bayesian approach, even though it has limitations such as the over-estimation of the tails of the null distribution of the FDR, for small sample sizes (Zhang(2007)). Future work should focus on the development of models for differential gene expression analysis that do not rely on the marginal distribution of the FDR. Our study was limited to the two microarray datasets from melanoma research, but we believe that the results would still hold when multiple datasets are considered.

Acknowledgments: We are indebted to Dr. William K. Kaufmann for the continuous and survival outcome data used in this study. This work was partially supported by a grant from the Simons Foundation (# 282714 to BO). LC was supported by a scholarship from the African Union.

References

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, **4(1)**:44 – 57.

Kaufmann, W. K., Nevis, K. R., Qu, P., Ibrahim, J. G., Zhou, T., Zhou, Y.,Sharpless, N. E. (2008). Defective Cell Cycle Checkpoint Functions in Melanoma Are Associated with Altered Patterns of Gene Expression. *Journal of Investigative Dermatology*, **128(1)**:175187.

Scharpf, Robert B. and Tjelmeland, Hkon and Parmigiani, Giovanni and Nobel, Andrew B. (2009). A Bayesian Model for Cross-Study Differential Gene Expression.*Journal of the American Statistical Association*, **5**: 158 – 176.

Lee, K. E., Sha, N., Dougherty, E. R., Vannucci, M., and Mallick, B. K. (2003). Gene selection: a Bayesian variable selection approach. *Bioinformatics*, **19(1)**:90 – 97.

Omolo, B., Carson, C., Chu, H., Zhou, Y., Simpson, D. A., Hesse, J. E., Kaufmann, W. K. (2013). A prognostic signature of G_2 checkpoint function in melanoma cell lines. *Cell Cycle*, **12(7)**: 1071 – 1082.

Simon, R., Lam, A., Li, M.-C., Ngan, M., Menenzes, S., and Zhao, Y. (2007). Analysis of gene expression data using BRB-Array Tools. *Cancer Informatics*,**3**: 11 – 17.

Smyth, G. K. (2005). Limma: linear models for microarray data. In:*Bioinformatics and computational biology solutions using R and Bioconductor*. R. Gentleman, V. Carey, S.Dudoit, R. Irizarry, W.Huber(eds.), Springer, New York, pp. 397 – 42.

Tibshirani, R.. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*. Series B (methodological), **58(1)**, 267 – 288.

Tusher, Virginia Goss and Tibshirani, Robert and Chu, Gilbert (2001).Significance analysis of microarrays applied to the ionizing radiation response. In: *Proceedings of the National Academy of Sciences* 5116 – 5121.

Winnepenninckx, V., Lazar, V., Michiels, S., Dessen, P., Stas, M., Alonso, S. R., Spatz, A. (2006). Gene Expression Profiling of Primary Cutaneous Melanoma and Clinical Outcome. *JNCI Journal of the National Cancer Institute*, **98(7)**: 472 – 482.

Zhang, S. (2007). A comprehensive evaluation of SAM, the SAM R-package and a simple modification to improve its performance. *BMC Bioinformatics*, **8(1)**: 230.

Gaussian Markov Random Fields within GAMLSS

Fernanda De Bastiani[1,4], Mikis D. Stasinopoulos[2], Robert A. Rigby[2], Miguel Uribe-Opazo[3] and Audrey H.M.A. Cysneiros[1]

[1] Statistics Department, Universidade Federal de Pernambuco, Recife/PE, Brazil
[2] STORM, London Metropolitan University, London N7 8DB, UK
[3] CCET, Universidade Estadual do Oeste do Paraná, Cascavel/PR, Brazil
[4] Postdoc in Statistics at Pontificia Universidade Catolica de Chile, Santiago, Chile

E-mail for correspondence: fernandadebastiani@gmail.com

Abstract: This paper describes the modelling and fitting of Gaussian Markov random field spatial components within a GAMLSS model. This allows modelling of any or all the parameters of the distribution for the response variable using explanatory variables and spatial effects. The response variable distribution is allowed to be a non-exponential family distribution.

Keywords: Discrete spatial analysis; Intrinsic autoregressive model; Random effects.

1 Introduction

Discrete spatial variation, where the variables are defined on discrete domains, such as regions, regular grids or lattices, can be modelled by Markov random fields (MRF). In statistics, Besag and Kooperberg (1995) considered the Gaussian intrinsic autoregressive model (IAR), a very important specific case of Gaussian MRF (GMRF) models. Wood (2006) presents IAR models within a generalized additive model (GAM) framework. Rigby et al. (2013) presented a simplified analysis of Munich rent data with very few covariates, modelling the μ parameter with a spatial effect using an IAR model term.
Section 2 discusses the GAMLSS framework. Section 3 discusses modelling and fitting of GMRF spatial components within GAMLSS models. Section 4 presents the infant mortality data set. Section 5 presents conclusions.

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

2 The GAMLSS framework

The distribution of the response variable is selected from a wide range of distributions available in the `gamlss` package in `R`, Rigby and Stasinopoulos (2005). This package includes distributions with up to four parameters, denoted by μ, σ, ν and τ. All the parameters of the response variable distribution can be modelled using parametric and/or nonparametric smooth functions of explanatory variables and/or random effect terms. A GAMLSS model assumes that, for $i = 1, \ldots, n$, independent observations Y_i have probability (density) function $f_Y(y_i|\theta^i)$ conditional on $\theta^i = (\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i})^\top = (\mu_i, \sigma_i, \nu_i, \tau_i)^\top$, each of which can be a function of the explanatory variables. In a random effects form it is given by:

$$\begin{aligned} g_1(\boldsymbol{\mu}) &= \boldsymbol{\eta}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \textstyle\sum_{j=1}^{J_1} \mathbf{Z}_{j1}\boldsymbol{\gamma}_{j1}, \\ g_2(\boldsymbol{\sigma}) &= \boldsymbol{\eta}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \textstyle\sum_{j=1}^{J_2} \mathbf{Z}_{j2}\boldsymbol{\gamma}_{j2}, \\ g_3(\boldsymbol{\nu}) &= \boldsymbol{\eta}_3 = \mathbf{X}_3\boldsymbol{\beta}_3 + \textstyle\sum_{j=1}^{J_3} \mathbf{Z}_{j3}\boldsymbol{\gamma}_{j3}, \\ g_4(\boldsymbol{\tau}) &= \boldsymbol{\eta}_4 = \mathbf{X}_4\boldsymbol{\beta}_4 + \textstyle\sum_{j=1}^{J_4} \mathbf{Z}_{j4}\boldsymbol{\gamma}_{j4}, \end{aligned}$$

where here the random effects parameters $\boldsymbol{\gamma}_{jk}$ are assumed to have independent (prior) normal distributions with $\boldsymbol{\gamma}_{jk} \sim N_{q_{jk}}(\mathbf{0}, \lambda_{jk}^{-1}\mathbf{G}_{jk}^{-1})$ and $\mathbf{G}_{jk}^{-1}$ is the (generalized) inverse of a $q_{jk} \times q_{jk}$ symmetric matrix $\mathbf{G}_{jk}$, where if $\mathbf{G}_{jk}$ is singular then $\boldsymbol{\gamma}_{jk}$ has an improper prior density function proportional to $\exp(-\frac{1}{2}\lambda_{jk}\boldsymbol{\gamma}_{jk}^\top\mathbf{G}_{jk}\boldsymbol{\gamma}_{jk})$.

3 Gaussian Markov Random Fields

A Markov random field (MRF) is a set of random variables having a Markov property based on conditional independence assumptions and described by an undirected graph, $\mathcal{G}$, where each vertex represents an areal unit and each edge connects two areal units and represents a neighbouring relationship, Rue and Held (2005).

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph (Whittaker, 2009) that consists of vertices $\mathcal{V} = (1, 2, \ldots, q)$, and a set of edges $\mathcal{E}$, where a typical edge is (m, t), $m, t \in \mathcal{V}$. Undirected is in the sense that (m, t) and (t, m) refer to the same edge. A random vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^\top$ is called a GMRF with respect to the graph $\mathcal{G}$, with mean $\boldsymbol{\mu}$ and precision matrix $\lambda\mathbf{G}$, if and only if its density has the form

$$\pi(\boldsymbol{\gamma}) \propto \exp\left[-\frac{1}{2}\lambda(\boldsymbol{\gamma} - \boldsymbol{\mu})^\top\mathbf{G}(\boldsymbol{\gamma} - \boldsymbol{\mu})\right] \qquad \text{and}$$

$$G_{mt} \neq 0 \iff (m, t) \in \mathcal{E} \text{ for } m \neq t,$$

where G_{mt} is the element of matrix $\mathbf{G}$ for row m and column t. It is denoted by $\boldsymbol{\gamma} \sim N(\boldsymbol{\mu}, \lambda^{-1}\mathbf{G}^{-1})$ where $\mathbf{G}^{-1}$ is the (generalized) inverse of $\mathbf{G}$.

When $\mathbf{G}$ is non-singular, another way to represent a GMRF, by its conditional mean and precision matrix, was given in Besag (1974), and known as the conditional autoregressive model (CAR). When $\mathbf{G}$ is singular the GMRF model can be represented by the IAR.

To incorporate IAR models within the GAMLSS model (1), set $\mathbf{Z}$ to be an index matrix defining which observation belongs to which area, and let $\boldsymbol{\gamma}$ be the vector of q spatial random effects and assume $\boldsymbol{\gamma} \sim N_q(0, \lambda^{-1}\mathbf{G}^{-1})$. In the following IAR model, the matrix $\mathbf{G}$ contains the information about the neighbours (adjacent regions), with elements given by $G_{mm} = n_m$ where n_m is the total number of adjacent regions to region m and $G_{mt} = -1$ if region m and t are adjacent, and zero otherwise, for $m = 1, \ldots, q$ and $t = 1, \ldots, q$. This model has the attractive property that conditional on λ and γ_t for all $t \neq m$, then $\gamma_m \sim N(\sum \gamma_t n_m^{-1}, (\lambda n_m)^{-1})$ where the summation is over all regions which are neighbours of region m.

4 The data set

The data set consists on the Parana infant mortality data (a region in Brazil) from 2010 with variables: `Infant Mortality`: number of infant deaths, `bornAlive`: number of children born alive, Firjan Index of city development, `illiteracy`: index of illiteracy, `GNP`: gross national product, `children-low_income (cli)`: proportion of children living in a household with half the basic salary, `population`: number of people, `Poor`: Proportion of individuals with household income per capita equal to or less than BRL 140.00 monthly, `fd`: factor for each city (provides the spatial explanatory variable). The `R` implementation of the IAR model as a predictor term for any parameter of the distribution of the response variable in a GAMLSS model is achieved by the `R` package `gamlss.spatial`. The beta binominal (BB) (Rigby et al, in press) final chosen fitted model is given by

$$\begin{array}{rcl} Y|N & \sim & BB(N, \hat{\mu}, \hat{\sigma}), \\ \text{logit}(\hat{\mu}) & = & -3.6281 + h_{11}(\log(Pop)) + 0.3039 \log(cli) + s(fd) \\ \log^*(\hat{\sigma}) & = & -9.879 + s(fd). \end{array}$$

where the h function is a smooth non-parametric function and s is an IAR spatial smoothing function and $log^*(\hat{\sigma}) = log^*(\hat{\sigma} - 0.00001)$ is specified by the link function `logshiftto0` in the `gamlss` to avoid the value of positive parameters reaching too close to zero. Figure 4 shows the fd effect on $logit(\hat{\mu})$ and the fd effect on $log^*(\hat{\sigma})$ where we can see that the infant mortality is higher in north and northwest region than the southeast region, and the variability is higher in north and southeast regions than in the west region of the Parana state.

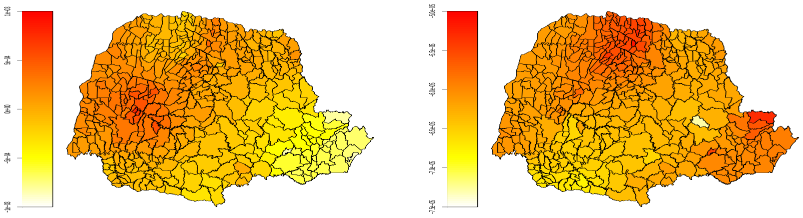


FIGURE 1. The fitted spatial effect for μ (left) and σ (right).

In the residual analysis we can see that the model fits well to the data, even though it is not perfect in the modeling of the variance in regions with low population.

5 Conclusion

The advantage of modelling spatial data within GAMLSS is that different distributions can be fitted and also it is possible, if needed, to model spatially any or all the parameters of the distribution.

Acknowledgments: The partial financial support from Fundação Araucária of Paraná State, Capes - 5925-13-4, CNPq and FACEPE from Brazil.

References

Besag, J. (1974) Spatial Interaction and the Statistical Analysis of Lattice Systems (with discussion), *Journal of the Royal Statistical Society, Series B*, **36**, 192 – 236.

Besag, J. and Kooperberg, C. (1995) On conditional and intrinsic autoregressions, *Biometrika*, **82**, 733 – 746.

Rigby, R. and Stasinopoulos, M. (2005) Generalized additive models for location, scale and shape (with discussion), *Applied Statistics*, **54**, 507 – 554.

Rigby, R.A., Stasinopoulos, D.M. and Voudouris, V. (2013) Discussion: A comparison of GAMLSS with quantile regression, *Statistical Modelling*, **13**,335 – 348.

Rigby, R.A., Statsinopoulos, D.M., Heller, G.Z. and Voudouris, V. (in press) *Distributions for Modelling Location, Scale, and Shape: Using GAMLSS in R*, Chapmand & Hall.

Rue, H. and Held, L. (2005) *Gaussian Markov Random Fields: Theory and Applications*, Chapman & Hall, New York.

Whittaker, J. (2009) *Graphical Models in Applied Multivariate Statistics*, Wiley, New York.

Wood, S. (2006) *Generalized Additive Models. An Introduction with R*, Chapman and Hall/CRC Press, Florida.

A Bayesian Test for Marginal Homogeneity in Contingency Tables

Helton Graziadei de Carvalho[1], Luis Gustavo Esteves[1]

[1] Institute of Mathematics and Statistics, University of Sao Paulo, Brazil

E-mail for correspondence: heltongc@ime.usp.br

Abstract: Matched sample studies have become popular in a wide range of applications especially those dealing with categorical variables. In this context, it is common to investigate if the marginal distributions are the same, the so-called Marginal Homogeneity (MH) hypothesis. Classical approaches to the problem of testing MH rely on the asymptotic (or approximate) distribution of the test statistics which may yield imprecise results in certain situations. To overcome these limitations, we develop the Full Bayesian Significance Test (FBST) for MH in two-dimensional contingency tables. The FBST is a procedure that has some important features such as: (i) it does not rely on asymptotic distributions (ii) it does not depend on the elimination of nuisance parameters (iii) it produces coherent results for nested hypotheses. Furthermore, we calculate p-values and compare them with the FBST. To summarize, we propose a coherent measure of evidence to test MH and compare it with classical approaches to the problem.,

Keywords: Marginal homogeneity; Categorical Data; Full Bayesian Significance Test.

1 Introduction

1.1 The Marginal Homogeneity Test

The problem of comparing the marginal discrete distributions for two paired-samples plays an important role in a variety of subjects such as: genetics, demography, politics and psychology (Agresti, 2002). We next present another example to illustrate the MH test as well as the techniques to be presented.

Example 1.1: Table 1 presents frequencies regarding the vision quality of the left and right eye for a group of 7477 women during the Second World War (Stuart, 1953). Let $\mathbb{C} = \{\text{Highest , Second, Third, Lowest}\}$ be the set of possible accuracies for each eye. In this context, considering a multinomial model we have

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

that the parameter space is: $\Theta = \theta = (\theta_{11}, \theta_{12}, \ldots, \theta_{14}, \ldots, \theta_{41}, \theta_{42}, \ldots, \theta_{44}) \in \mathbb{R}_+^{16}, \sum_{i=1}^4 \sum_{j=1}^4 \theta_{ij} = 1\}$, where θ_{ij} is the probability that the individual's right eye vision is classified in the j-th category of $\mathbb{C}$ and the left eye quality is classified in the i-th category of $\mathbb{C}$, $\forall i, j = 1, \ldots, 4$. The sample space is: $\chi = \{(n_{11}, n_{12}, n_{13}, n_{14}, \ldots, n_{41}, \ldots, n_{44}), \in \mathbb{N}^{16} : \sum_{i=1}^4 \sum_{j=1}^4 n_{ij} = 7477\}$, where n_{ij} represents the count of cell (i, j), $i, j = 1, \ldots, 4$.

TABLE 1. Unaided distance vision of 7455 women in Britain (Stuart, 1953)

	Highest	Second	Third	Lowest	Total
Highest	1520	266	124	66	1976
Second	234	1512	432	78	2256
Third	117	362	1772	205	2456
Lowest	36	82	179	492	789
Total	1907	2222	2507	841	7477

In this scenario, one hypothesis of interest is $H : \theta_{i+} = \theta_{+i}, i = 1, 2, 3$, where $\theta_{i+} = \sum_{j=1}^k \theta_{ij}$ and $\theta_{+i} = \sum_{j=1}^k \theta_{ji}$, that is, H means that individuals have the same individuals have the same quality in both eyes (marginal homogeneity). A further hypothesis that is common to be investigated in matched-sample studies is $H' : \theta_{ij} = \theta_{ji}, (i, j) \in \mathbb{C} \times \mathbb{C}$ such that $i \neq j$, which is called symmetry. The p-value obtained from the likelihood ratio test for H (Madansky, 1956) is equal to 0.009. Additionally, using a generalization of McNemar test (Bowker, 1948), the p-value for H' is 0.080. Hence, adopting a 5% (or even 1%) significance level, we conclude simultaneously that the distributions of the qualities of vision are different, while the corresponding joint distribution is symmetric, which seem to be inconsistent because $H' \subset H$ (Agresti, 2002).

1.2 The Full Bayesian Significance Test (FBST)

The FBST (Pereira and Stern, 1999) was developed as an alternative to overcome some difficulties usually met by frequentist and bayesian tests. Suppose a bayesian statistical model, i.e., $\Theta \subset \mathbb{R}^k$ is the parameter space and $\chi \subset \mathbb{R}^k$ is the sample space. Also, $f(\theta)$ is a prior probability density over Θ and $L_x(\theta)$ is the likelihood function generated by an observation $x \in \chi$. Consider that a sharp hypothesis $H : \theta \in \Theta_0$ (that is, $dim(\Theta_0) < dim(\Theta)$) is to be tested. The FBST is based on the measure of evidence, called e-value, described in the sequel. To calculate the e-value, let $f(\theta|x)$ be the posterior density function for θ given by

$$f(\theta|x) \propto f(\theta) L_x(\theta).$$

Let $T_x = \{\theta \in \Theta : f(\theta|x) > \sup_{\theta \in \Theta_0} f(\theta|x)\}$ be the tangential (to Θ_0) set which is composed of the points in parameter space that are more consistent with x than the posterior mode under the null hypothesis. The e-value in favor of H is defined as

$$ev(\Theta_0; x) = 1 - P(\theta \in T_x|x).$$

As defined by Pereira and Stern (1999), the FBST is the procedure that rejects H whenever $ev(\Theta_0; x)$ is small. In addition, it should be emphasized that the posterior density is the only requirement to calculate e-values. In the next section, we develop the FBST for MH in two-way contingency tables.

2 The FBST for Marginal Homogeneity

Suppose a random vector $X|\theta \sim$ Multinomial(n,θ), $\theta = \{\theta = (\theta_{11}, \dots \theta_{kk}) \in \mathbb{R}_+^{k^2}, \sum_{i=1}^{k}\sum_{j=1}^{k}\theta_{ij} = 1\}$ and the sample space is $\chi = \{(n_{11}, \dots, n_{kk}) \in \mathbb{N}^{k\times k} : n_{++} = n\}$. The marginal homogeneity hypothesis is written as

$$H : \theta_{i+} = \theta_{+i}, i = 1, \dots, k-1.$$

Suppose that $\theta \sim$ Dirichlet(a). Then, by using Bayes' Theorem, we have that the kernel of the posterior density is

$$f(\theta|x) \propto \left[\prod_{i=1}^{k}\prod_{j=1}^{k}\theta_{ij}^{n_{ij}+a_{ij}-1}\right]1_{\Theta}(\theta).$$

Thus, $\theta|x' \sim$ Dirichlet(x'), where $x' = (n_{11}+a_{11}, n_{12}+a_{12}, \dots, n_{1k}+a_{1k}, \dots, n_{k1}+a_{k1}, \dots, n_{kk}+a_{kk})$.To obtain comparisons with frequentists solutions, we define $a_{ij} = 1$, $\forall\ i, j \in \{1, \dots, k\}$. To calculate the e-value in favor of the marginal homogeneity hypothesis for two-dimensional contingency tables, it is necessary to specify the tangential set T_x for Θ_0. In order to do so, we first need to maximize the kernel of the log-posterior density that is

$$\text{Maximize} \sum_{i=1}^{k}\sum_{j=1}^{k}(n_{ij} + a_{ij} - 1)\log\theta_{ij}$$

subject to k constraints $\theta_{i+} = \theta_{+i}, \quad i = 1, \dots, k-1$ and $\sum_{i=1}^{k}\sum_{j=1}^{k}\theta_{ij} = 1$. Using a vector of Lagrange Multipliers $\lambda = (\lambda_0, \lambda_1, \dots, \lambda_{k-1})$, we need to maximize

$$\mathcal{L}(\theta, \lambda) = \sum_{i=1}^{k}\sum_{j=1}^{k}(n_{ij} + a_{ij} - 1)\log\theta_{ij} - \lambda_0\left(\sum_{i=1}^{k}\sum_{j=1}^{k}\theta_{ij} - 1\right) - \sum_{l=1}^{k-1}\lambda_l(\theta_{l.} - \theta_{.l}).$$

It is easy to show that $\lambda_0 = n$, $\tilde{\theta}_{ii} = \frac{n_{ii}+a_{ii}-1}{n}$, $i = 1, \dots, k$, which is equal to the corresponding coordinate of the posterior mode, and $\tilde{\theta}_{ij} = \frac{n_{ij}+a_{ij}-1}{n+\lambda_i-\lambda_j}$
From Equation 6, it is possible to obtain the Lagrange Multipliers regardless of θ. Next, we use Equation 5 to determine the estimator of θ under H.
After finding the maximum of the posterior density under H, we need to calculate the posterior probability of the tangential set, $\mathbb{P}(\theta \in T_x|x)$. We perform this calculation by means of Monte Carlo method: we generate $\theta_1, \theta_2, \dots, \theta_M$ of the posterior density and compare their densities with the maximum density under H. Let $\theta^* = \arg\max\{\pi(\theta|x) : \theta \in \Theta_0\}$ and define

$$1_A(\theta_i) = \begin{cases} 1, & \text{if} \quad f(\theta_i|x) \geq f(\theta^*|x), \\ 0, & \quad\text{otherwise,} \end{cases}$$

$i = 1, \ldots, M$.
Thus, we approximate the evidence $ev(\Theta_0; x)$ in favor of the marginal homogeneity hypothesis for a $k \times k$ contingency table using the Monte Carlo Method, that is,

$$\tilde{ev}(\Theta_0; x) = 1 - \frac{\sum_{i=1}^{M} 1_A(\theta_i)}{M}.$$

In Example 1.1, the evidence for marginal homogeneity is $ev(\Theta_0; x) = 0.68$ whereas for symmetry is $ev(\Theta_0'; x) = 0.20$, leading us to conclude that the data give more evidence to support MH hypothesis than the Symmetry hypothesis. Note that the generalization for k-way contingency tables is trivial.

References

Agresti, A. (2002). Categorical Data Analysis. *John Wiley* & *Sons.*

Bowker, A.H. (1948). A Test for Symmetry in Contingency Tables. *Journal of the American Statistical Association.*

Madansky, A. (1963). Tests of Homogeneity for Correlated Samples. *Journal of the American Statistical Association*

Stuart, A. (1953). The Estimation and Comparison of Strengths of Association in Contingency Tables. *Biometrika.*

Pereira, C.A.B. and Stern, J.M. (1999). Evidence and Credibility: Full , Bayesian Significance Test for Precise Hypothesis. *Entropy.*

Markov transition models for ordinal responses applied to pig behavior

Idemauro Antonio Rodrigues de Lara[1][2], John Hinde[2], Ariane Cristina de Castro[1]

[1] Luiz de Queiroz College of Agriculture, University of São Paulo, Brazil.
[2] School of the Mathematics, Statistics and Applied Mathematics, National Univerisity of Ireland, Galway (NUIG).

E-mail for correspondence: idemauro@usp.br

Abstract: In this work, the data concerns the severity of lesions related to the behaviour of pigs. The experimental design corresponded to two levels of environmental enrichment and four levels of genetic lineages in a completely randomized 2×4 factorial. The data was collected longitudinally over four time occasions. We consider the use of transition models for analysing these data. The methodology allows for the choice of a model that can be used to explain the relationship between the severity of lesions in the pigs and the use of environmental enrichment.

Keywords: ordinal longitudinal data, animal behavior, transition probabilities.

1 Introduction

There is much interest in studies related to the behavior of animals in various areas such as cattle, goat, pig, and fish farming driven by concern for animal welfare and for improved production and reproductive value. The development of models and statistical methods in this area is also an object of interest. An intrinsic characteristic of the response data measured in these studies is that they are on a nominal or ordinal scale (categorical data).

Another characteristic inherent in these studies is that they are longitudinal, so there is the need to consider a possible correlation between the observations made on the same animals. Two model classes commonly used for longitudinal data analysis are marginal models (Liang and Zeger, 1986) and random effects models (Diggle et al. 2002, Molenberghs and Verbeke, 2005). However, there are situations where it is likely that the state of an individual on the previous occasion influences the individual's current state. The interest is in what

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

happens to the categorical responses from one moment to another and to assess the possible effects of covariates. To meet this goal, we consider the use of Markov transition models (Ware, Lipsitz and Speizer 1988, Lindsey 2004). These models are based on stochastic processes. Here, we consider a discrete time process with discrete state space and a first-order Markov assumption, that is the transition probabilities at time t only depend upon the current state with $\pi_{ab}(t-1,t) = P(Y_t = b \mid Y_{t-1} = a)$, with $a, b \in S = \{1, 2, \ldots, k\}$ and $t \in \tau = \{0, 1, \ldots, T\}$. To simplify the notation, we write $\pi_{ab}(t-1,t) = \pi_{ab}(t)$ for the transition probability at time t from state a to state b. This work presents an extension of the proportional odds model for this transition model setting.

2 Material

The data are the result of research conducted by Castro (2015), during the months of March to July 2014 at a commercial farm group in Brazil. The treatment structure was in a 2×4 factorial, with factors: Environmental enrichment levels: E1 with and E2 without; Genetic lineage levels: L1 a synthetic line; L2 from crossing two distinct lineages (Pietrain); L3 from the Landrace line; and L4 from the Large White line. In all, 124 animals were used across the eight treatment combinations. Each treatment combination consisted of a pen with 16 animals, with the animal being considered as the experimental unit. The response variable of interest is a score measuring lesions on the front of the animal, that were classified as follows: 0 an absence of lesions; 1 a moderate degree of lesions; and 2 for serious lesions. Four monthly evaluations were made over the duration of the study.

3 Methods

We consider the proportional odds model (McCullagh, 1980) for the ordinal response and incorporate the longitudinal dependence using a Markov chain model by including the response category at the preceding time as a covariate. In this context, $\mathbf{x}_{it} = (x_{it1}, x_{it2}, \ldots, x_{itp}, x_{it(p+1)})'$ is the vector of $(p+1)$ covariates associated with the ith individual at the tth transition, where x_{it1} denotes the previous state. The model is:

$$\eta = \log\left(\frac{\gamma_{ab}(t)(\mathbf{x})}{1-\gamma_{ab}(t)(\mathbf{x})}\right) = \lambda_{abt} + \delta_t'\mathbf{x}$$

where λ_{abt} is an intercept and $\delta_t' = (\alpha_t, \beta_{1t}, \ldots, \beta_{p,t})$ is a vector of unknown parameters. The transition cumulative probabilities are specified by:

$$\gamma_{ab}(t)(\mathbf{x}) = \frac{\exp(\lambda_{ab(t)} - \delta_t'\mathbf{x})}{1+\exp(\lambda_{ab(t)} - \delta_t'\mathbf{x})} \qquad\qquad b = 1, 2, \ldots, k-1,$$

where $\gamma_{ab}(t)(\mathbf{x}) = \mathrm{P}(Y_{jt} \le b \mid \mathbf{x}) = \pi_{ab}(t)(\mathbf{x}) + \ldots + \pi_{ab}(t)(\mathbf{x}),\;\; b = 1, \ldots, k-1$. Here, with just four time occasions there are only three transitions of order one and so it is not sensible to consider higher order chains. In addition, for simplicity, we also assume stationarity. Several models are considered; Model 1: all main effects and interaction between lineage and enrichment; Model 2: all main effects and no interaction; Model 3: no lineage effect; Model 4: only the previous response covariate; Model 5: interaction between enrichment and previous response. For example, the linear predictor for model 5 is:

$$\begin{aligned}\eta_{lts} &= \lambda_{ab} - [\beta_e \text{ enrichment}_e + \beta_s \text{previous response}_s + \\ &+ \beta_{es}\text{enrichment} * \text{previous response}_{es}]\end{aligned}$$

with $l = 1, 2, 3, 4$; $e = 1, 2$; and $s = 0, 1, 2$. For fitting these transition models we used the package `ordinal` (Christensen, 2011) available in `R` software (R Development Core Team). Model selection based on likelihood ratio tests leads us to choose Model 5.

4 Results

Table 1 shows the parameter estimates for the selected Model 5, in particular we see the significance of previous response and how much it may influence the effect of the enrichment covariate.

TABLE 1. Parameter estimates and Standard errors (s.e.) for the selected transition Model 5.

Parameters	**Estimates**	**s.e.**	**p-value**
λ_{a1}	0.2522	0.2520	
λ_{a2}	2.7404	0.2937	
enrichment(E2)	0.6341	0.4035	0.1160
previous response(1)	0.8617	0.3232	0.0076
previous response(2)	0.7408	0.4283	0.0837
previous(1):enrichment(E2)	−0.6451	0.4987	0.1957
previous(2):enrichment(E2)	0.8144	0.5815	0.1613

Table 2 gives the fitted transition probabilities from Model 5 and shows that given that an animal is in the good condition (state 0), with environmental enrichment, it has probability 0.5627 to continue in the same condition, whereas if it has no environmental enrichment, this probability falls to 0.4056. Whereas if the precondition of the animal is bad (state 2), it has probability 0.3802 to change to the good state, while without environmental enrichment this probability falls to 0.1259.

It was found that the use of environmental enrichment is beneficial and gives some degree of animal protection, in that if an animal has good or severe lesions then the probability that it will move to a better state is, in general, higher than for the animals receiving no enrichment.

Acknowledgments: Special Thanks to the funding agency FAPESP (São Paulo state, Brazil)

TABLE 2. Estimates of the transition probabilities by Model 5.

		Enrichment					
		E1			E2		
Future Response		0	1	2	0	1	2
Previous	0	0.5627	0.3766	0.0606	0.4056	0.4858	0.1084
Response	1	0.3521	0.5153	0.1325	0.3546	0.5140	0.1312
	2	0.3802	0.5005	0.1192	0.1259	0.5084	0.3656

References

Castro, A.C. (2015) *Comportamento e desempenho sexual de suìnos reprodutores em ambientes enriquecidos.* Phd Thesis: University of São Paulo.

Christensen, H. (2011) *Analysis of ordinal data with cumulative link models estimation with the R-package ordinal.* See `www.R-project.org`.

Diggle, P.J., Heagerty, P., Liang K-Y and Zeger, S.L. (2002) *Analysis of longitudinal data.* Oxford: University Press.

Lindsey, J.K. (2004) *Statistical analysis of stochastic processes in time.* New York: Cambridge University Press.

McCullagh, P. (1980). Regression Methods for Ordinal Data. *Journal of The Royal Statistical Society*, **42**, 109 – 142.

Molenberghs, G. and Verbeke, G. (2005) *Models for discrete longitudinal data.* New York: Spriger-Verlag.

R Development Core Team *A language and environment for statistical computing 3.2.* See `www.R-project.org`.

Ware, J. H., Lipsitz, S. and Speizer, F. E. (1988) Issues in the Analysis of Repeated Categorical Outcomes. *Statistics in Medicine*, Chichester, **7**, 95 – 107.

Describing and understanding functioning in Spinal Cord Injury - a graphical model approach

Cristina Ehrmann-Bostan[1,2], Birgit Prodinger[1,2], Gerold Stucki[1,2]

[1] Swiss Paraplegic Research (SPF), Nottwil, Switzerland
[2] Department of Health Sciences and Health Policy, University of Lucerne, Switzerland

E-mail for correspondence: `cristina.bostan@paraplegie.ch`

Abstract:
Background:The understanding of functioning in SCI is necessary for planning and organization of health services provisioning, social policy and rehabilitation management. The SwiSCI Cohort Study aimed to respond to this necessity by collecting the most relevant information on functioning of people with SCI living in Switzerland.
Objective: To describe and understand functioning in people with SCI living in Switzerland.
Methods: Data from the Swiss Cohort Study Community Survey was used. Firstly, descriptive statistics were used to summarize the sample characteristics. Secondly, a univariate analysis was considered for calculating the prevalence of relevant problem in each functioning aspect. Thirdly, graphical model was applied to visualize the association structure.
Results: Overall, 1549 persons participated in the Survey, with 71.5 % male and median age of 52 years. Approximatively 69 % had paraplegia and 58 % incomplete lesions. The functioning areas where more than 60 % persons reported problems or limitations were: sexual functions, spasticity, chronic pain, bladder dysfunction, bowel dysfunction, tireness, stairs, doing housework, sports, activities outdoors. Mental health, transfer, washing and dressing are connected components shown when visualizing functioning in people with SCI.
Conclusions: Graphical models can be used to describe and understand functioning in people with SCI.

Keywords: Spinal Cord Injury; Prevalence; Graphical Model; Functioning

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

1 Materials and Methods

1.1 Subject Characteristics

Data from the Swiss Cohort Study Community Survey was used. The source of subjects for this study was the Swiss Paraplegic Association, 3 specialized SCI rehabilitation centres, and a SCI-specific home care institution (Brinkhof, 2016).

1.2 Measures

Demographic information was collected within the Starting Module of this survey: gender, age, SCI aetiology (traumatic and non-traumatic), and lesion group (paraplegia and tetraplegia).

In the Basic Module Survey respondents were asked about the presence and severity of difficulties/problems in functioning in their everyday life. Reliable and valid instruments were used to operationalize functioning: SCI Secondary Conditions Scale,Self-Administered Comorbidity Questionnaire, 36-item Short Form, SCI Independence Measure Self-Report, Utrecht Scale for Evaluation Rehabilitation-Participation.

1.3 Statistical Analysis

Firstly, descriptive statistics were used to summarize the sample characteristics. Secondly, a univariate analysis was considered for calculating the prevalence of relevant problem in each functioning aspect. For this purpose, the dichotomization strategy was applied for each question, where 0 = non existing or insignificant and 1 = existing and relevant problem/limitation . Thirdly, the undirected graph model, called skeleton, was estimated using the PC algorithm implemented by Kalisch et al for visualizing the association structure (Kalisch, M et al, 2007).

2 Results

Table 1 shows descriptive characteristics of the population.

Figure 1 shows that sexual functions, spasticity, chronic pain, bladder dysfunction, bowel dysfunction, tireness, stairs, doing housework, sports, activities outdoors, are the most ofen reported functioning areas with problems or limitations.

In the Figure 2 the association structure between aspects of functioning is presented. Mental health, transfer, washing and dressing are connected components shown when visualizing functioning in people with SCI.

TABLE 1. Characteristics of study participants (n=1549)

Characteristic	N	%
Male	1107	71.5
Female	442	28.5
Age (years, median)	52	
Paraplegia, incomplete	577	37.5
Paraplegia, complete	486	31.6
Tetraplegia, incomplete	314	20.4
Tetraplegia, complete	160	10.4
Traumatic	1202	78.4
Non-traumatic	332	21.6

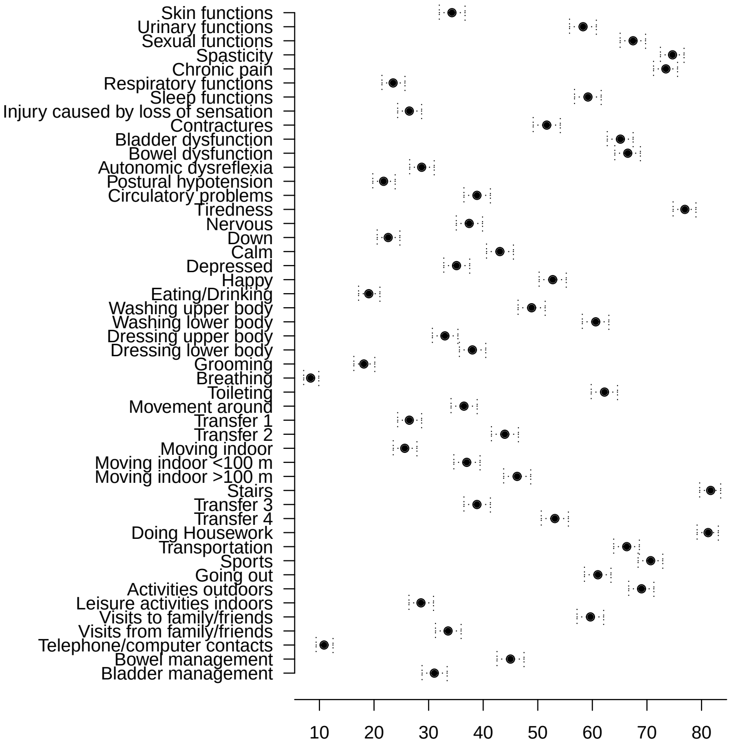


FIGURE 1. Prevalence and 95% confidence interval of reported problems or limitations in various functioning areas.

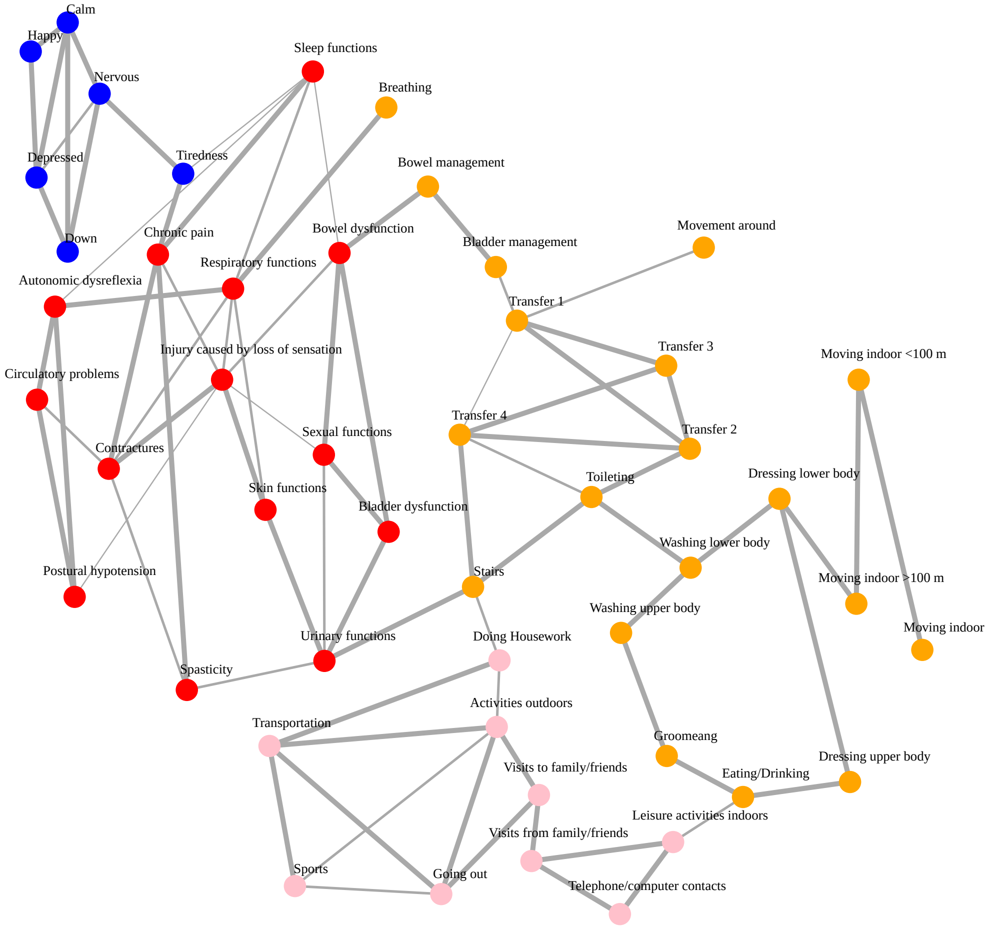


FIGURE 2. Association between aspects of functioning in SCI population identified across 50 random samples. The edge thickness is proportional to edge stability. Edges identified in less than 25 random samples were deleted.

References

Brinkhof, MW (2016). Basic description of the Swiss Spinal Cord Injury Cohort Study (SwiSCI). *Journal of Rehabilitation Medicine*, **48**.

Kalisch, M and Bhlmann, P (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, **8**, 613 – 636.

Econometric detection of speculative bubbles: application of the Markov regime switching model

Mouloud Elhafidi[1], Abdessamad Ouchen[2]

[1] University Sid Mohamed Ben Abdellah, Morocco
[2] University Sid Mohamed Ben Abdellah, Morocco

E-mail for correspondence: ouchenencg@gmail.com

Abstract: The choice of our topic is due to the recurrence of financial crises. The world today is deeply unstable and subject to uncertainties and big surprises. Finance is known for two regimes: state of stability and state of crisis. Therefore, in order to understand the cyclical asymmetries in the series of yields of the main indices of the world, one has to resort to non-linear specifications that distinguish between upswings and downturns. We estimated a switching model in both states and with a specification autoregressive of order 1, the monthly first difference of the S&P 500 during the period running from December 1999 to December 2015. This model allowed us to confirm the existence of two regimes distinct on the Wall Street Stock Exchange, namely the state of crisis and that of stability. It allowed the detection of three bubbles: the dot.com bubble (1998-2000), the real estate bubble (1995-2006) and the Chinese financial bubble (2014-2015). Indeed, the probability of being in crisis phase (probability smoothing) is greater than 0.6 after the crisis of TMT (2000-2001), after the financial crisis between 2007 and 2008, after the European debt crisis in 2010 and after the Chinese financial crisis in 2015.

Keywords: Markov regime switching model; Speculative bubbles.

1 Introduction

During the last twenty years, no fewer than ten financial crises: the collapse of Barings Bank in 1995, the Mexican crisis between 1994 and 1995, the Thai crisis between 1997 and 1998, the Russian crisis in 1998, the near collapse of LTCM in 1998, bursting the Internet bubble between 2000 and 2001, the Argentine crisis between 2001 and 2002, the financial crisis between 2007 and 2008, the sovereign debt crisis in the euro area began at the end of 2009 and the bursting of the

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

Chinese financial bubble in 2015. This litany of financial crises shows indeed that our universe became turbulent and marked by the unthinkable.
The world today is deeply unstable and subject to uncertainties and big surprises. Finance is known for two regimes : state of stability and state of crisis. Therefore, in order to understand the cyclical asymmetries in the series of yields of the main indices of the world, one has to resort to non-linear specifications that distinguish between upswings and downturns.
Having enjoyed success in the analysis of quarterly gross domestic product of the United States, the Markov switching model (Hamilton(1989)) constitutes adequate econometric tool to be taken into consideration for cyclical asymmetries in the series of our variable, namely: the monthly first difference of the S&P 500, that is to say the nonlinearity in this serie. This is an approach which can identify and detect turning points in both peaceful and crisis phases of financial time series.
In this article, we focus on the detection of financial shocks and speculative bubbles by the Markov switching model.

2 Application of the Markov regime switching model

Inspired by Hamilton (1989), we will estimate a switching model in both states and with a specification autoregressive of order 1, the monthly first difference of the S&P 500 during the period running from December 1999 to December 2015.

TABLE 1. **Estimation of the Markov regime switching model**.

Coefficient	Estimated coefficient of Markov regime switching model
μ_0	21.80178
	[0.0000]
μ_1	-75.87128
	[0.0000]
β	-0.198561
	[0.0068]
$\ln(\sigma)$	3.696109
	[0.0000]

Where: [.] is critical probability;

$$DS\&P500_t = \mu_0(1-S_t) + \mu_1 S_t + \beta(DS\&P500_{t-1} - \mu_0(1-S_{t-1}) - \mu_1 S_{t-1}) + \varepsilon_t$$

with: $\varepsilon_t \sim N(0;\sigma^2)$; $S_t \in \{0;1\}$ where the state $S_t = 1$ is the crisis regime and the state $S_t = 0$ is the stability regime; $P_{ij} = Pr(S_t = j/S_{t-1} = i)$ is the probability of moving from state i to state j and $\sum_{j=0}^{1} P_{ij} = 1$ for $i \in \{0;1\}$; that is to say: $P_{01} = 1 - P_{00}$ and $P_{10} = 1 - P_{11}$; and $P_{00} = 0,9146$ and $P_{10} = 0,3871$.

After the analysis of this model, the first outcome is that the estimated parameters are significant at the 5% statistical; and, secondly, that there are changes in different schemes in the first difference of the S & P 500. In fact, there are two states : an optimistic or stable state, positive average equal to 21.80, and another pessimistic or crisis, negative mean of - 75.87. In addition, the state of stability, which has a transition probability of $P_{00} = 0.9146$, is more persistent compared to the crisis, which has a transition probability of $P_{11} = 0.6129$. Moreover, the unconditional probabilities of the state of stability and the crisis, which are equal to $\pi_0 = \frac{1-P_{11}}{2-P_{11}-P_{00}} = 0.8193$, $\pi_1 = \frac{1-P_{00}}{2-P_{11}-P_{00}} = 0.1807$ respectively and indicate that, for a given sample of the first difference of the index of S & P 500 close to 18.07 % of the observations should be in a state of crisis. It was also found that the conditional expected duration in the state of crisis equals 2.58 months. That is to say, we can expect, on average, a high volatility period lasts about two and a half month. It should also be noted that Wall Street Stock Exchange has a chance to move from the state of crisis in t-1 to the state of stability ($P_{10} = 0.3871$) greater than the chance of moving from the state of stability in t-1 to the state of crisis ($P_{01} = 0.0854$).

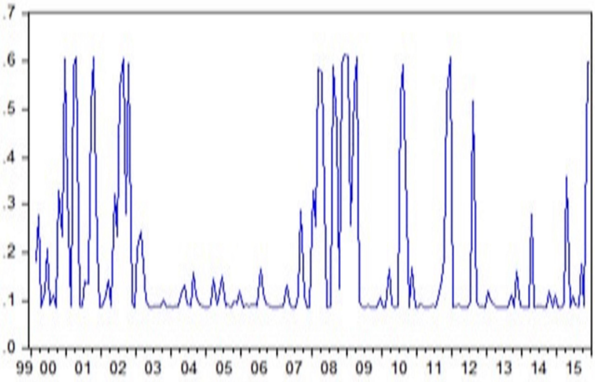


FIGURE 1. **Smoothed probability of the crisis regime**

Where : $Pr(S_t = 1|DS\&P500_t; \theta)$ is the smoothed probability of the crisis regime; and $\theta = (\mu_0, \mu_1, \beta, \sigma^2, \pi_0, \pi_1)'$ is the vector of parameters to be estimated.

Given the evolution of the probability of being in crisis phase in the Wall Street Stock Exchange (Figure 1), we can detect that this probability is greater than or equal to 0.6 for high volatile periods of the first difference of the S & P 500, after the crisis of TMT between 2000 and 2001, after the financial crisis between 2007 and 2008, after the European debt crisis in 2010 and after the financial crisis in China in 2015.

3 Conclusion

All financial crises which occur between the periods 2000 and 2015 were detected by the Markov switching model. Our model has allowed the detection of four financial crises in 2001, 2008, 2010 and 2015, respectively, after the bursting

of the Internet bubble (1998-2000), the bursting of the housing bubble (1995-2006),the European debt crisis in 2010, and the bursting of the Chinese financial bubble in 2015. Moreover, one can expect, on average, a high volatility period lasts about two and a half month. It should also be noted that Wall Street Stock Exchange has a chance to move from the state of crisis in t-1 to the state of stability (38.71%) greater than the chance of moving from the state of stability in t-1 to the state of crisis (8.54%).
Certainly, knowledge of financial shocks that may occur allows taking proactive measures to ensure financial stability. However, their forecasts are very delicate. Evidenced, questioning "why no one did not see this crisis coming?", and sending in November 2008 by Queen Elizabeth to professors from the famous University London School of Economics. In addition, Isaac Newton, part of investors ruined when the bursting of the bubble of the Companion of the South Seas in 1720, confesses: "I can measure the motions of heavenly bodies, but I cannot measure human nonsense".

Acknowledgments: Special Thanks to the Organizing and Scientific Committees.

References

Aglietta, M., and Rigot, S. (2009). *Crise et rénovation de la finance.* Odile Jacob.

Ben Hammouda, H., et al. (2010). *Crise... Naufrage des économistes ?.* Groupe De Boeck s.a.

Bourbonnais, R. (2009). *Econométrie.* Dunod.

De Boissieu, C. (2009). *Les Systèmes financiers : mutations, crises et régulation.* Economica.

Evans, G. (1991). Pitfalls in testing for Explosive Bubbles in Asset Prices. In: *American Economic Review,***81**, 922 – 930.

Froot, K. A., and Obstfeld, M. (1991). Intrinsic Bubbles: The Case of Stock Prices. In:*American Economic Review*, **81**, 1189 – 1214.

Hamilton, J. D. (1989). A New Approach to the economic analysis of non-stationary time series and the business cycle. In:*Econometrica,* **57**.

Lacoste, O. (2009). *Comprendre les crises financières.* Groupe Eyrolles.

Use of GAM for determining the efficacy of RIRS in kidney stone removal

Nazia P. Gill[1], Letterio D'Arrigo[2], Francesco Savoca[2], Angela Costa[2], Astrid Bonaccorsi[2], Ernst C. Wit[1] and Michele Pennisi[2]

[1] Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, Groningen, the Netherlands
[2] Urology Unit, Cannizzaro Hospital, Catania, Italy

E-mail for correspondence: n.p.gill@rug.nl

Abstract: In the last years the introduction of endoscopic intrarenal surgery (RIRS) has reduced the need for percutaneous treatments. Moreover, it has been suggested as an alternative treatment for patients with stones larger than 2 cm in diameter. The aim of this study is to identify which type the patients can reap the benefits from RIRS, as well as those that are better off by a percutaneous procedure, in terms of a stone free outcome of the procedure, postoperative complications, such as sepsis, and operative time. A total of 106 patients with renal calculi were treated with RIRS. The overall stone free rate was 77%. In patients with stone diameters below 2 cm the stone free rate was 85%, which decreased to 55% for stones over 2 cm. The infundibular length (p-value 0.186), width (p-value 0.2074), angle (p-value 0.252), volume (p-value 0.3573) and stone density (p-value 0.7784) did not correlate with overall stone free status. Considering only patients with stones in the lower calyx, smaller infundibular angles negatively influenced the stone free rate (p-value 0.001). Operation time was less in the stone free patients (p-value 0.0003) and a positive correlation with stone size (p-value 0.0002) was found. In patients with DJ the incidence of sepsis was 57%, which was 24% for those with nephrostomy and 14% in all others. We conclude that the factors that influenced stone free rate of RIRS are stone diameter, the number of stones and, when urinary stones are present in the lower calyx. The effect of stone density on the stone free probability is non-monotone.

Keywords: Generalized additive modelling, retrospective clinical trial, RIRS, stone free, sepsis.

[1] *joint first authors

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

1 Introduction

Retrograde intrarenal surgery (RIRS) with holmium laser lithotripsy is widely considered an alternative to percutaneous litholapaxy (PCNL) for patients with renal stones that are resistant to extracorporeal shockwave lithotripsy (ESWL). For such patients with lower calyx stones smaller than 1 cm retrograde surgery can be considered a first choice treatment with an effectiveness similar to PCNL, but with a lower comorbidity. Whether RIRS can be considered an alternative to percutaneous treatment even in the presence of renal stones between 1-2 cm is a matter of discussion, as international guidelines do not express a preference between two procedures (Turk et al. 2014).

In this study we are evaluating the effect of the number of stones, their location, maximum diameter, volume, area, stone density and calyceal anatomy (infundibular length, width and angle) to identify the parameters that are predictive of stone free rate (SFR) in a RIRS intervention. The aim of the study is to identify the relevant patient features to see who can reap the benefits from RIRS and who might be better off by the traditional percutaneous procedure. In addition, the influence of these parameters on operative time and the onset of postoperative complications with particular attention to sepsis are analyzed.

2 Data and Statistical Analysis

A retrospective analysis of 106 patients who underwent RIRS interventions (single or multiple) between March 2011 to December 2013 for the treatment of kidney stones in urology unit of Cannizzaro Hospital in Catania was performed. The statistical analysis was performed using the statistical software package R (version 3.0.3). P-values smaller than 0.05 were considered statistically significant. To assess the relationship between the stone free (SF) rate and measures of stone burden (i.e. number of stones, stone density and stone diameter) a generalized additive model with a logistic link function was used. The final model was selected using Generalized Cross Validation (GCV) for obtaining a robust predictive model.

$$\operatorname{logit}\left[P_{SF}\left(x\right)\right] = S_1\left(x_1, x_2\right) + S_2\left(x_3\right) \tag{1}$$

where

$$\begin{aligned} S_1\left(x_1, x_2\right) &= \sum_{j=1}^{k_1}\sum_{l=1}^{k_2} b_j\left(x_1\right) b_l\left(x_2\right)\beta_{jl}, \\ & \qquad\qquad k_1, k_2 = 5 \\ S_2\left(x_3\right) &= \sum_{i=1}^{k_3} b_i\left(x_3\right)\beta_i, \\ & \qquad\qquad k_3 = 10 \end{aligned}$$

3 Results

3.1 Joint Analysis of stone burden on stone free rate

The joint analysis reveals that effect is highly non-linear (figure 1(b)).

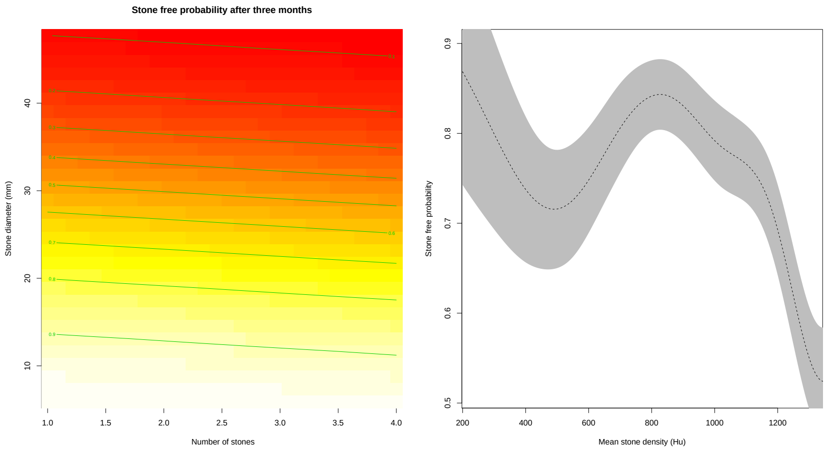


FIGURE 1. (a) joint effect of stone diameter and number of stones on RIRS success (b) Dependence of stone free probability after three months on mean stone density: for a patient with 2 stones and stone diameter of 2 cm.

3.2 Factors affecting RIRS complications

Table 1: Prediction table of sepsis incidence based on DJ status, age and gender.

		Age			
DJ status	Sex	20	40	60	80
DJ	M	42%	23%	10%	4%
	F	64%	41%	22%	10%
non-DJ	M	27%	13%	6%	2%
	F	47%	26%	12%	5%

4 Discussion

Intrarenal retrograde laser lithotripsy and SWL are considered the treatment of choice in patients with renal stones with a diameter smaller than 1 cm, whereas EAU guidelines recommend percutaneous procedures in patients with stones over 2 cm. The guidelines are impartial when it comes to stones between 1 and 2 cm, if there are no further factors. However, in the presence of a lower pole stone between 1 and 2 cm the endourological treatment is recommended as treatment of first choice, as the SWL is less effective in those circumstances. The reason is the high percentage of residual fragments after a SWL intervention, which can be exacerbated by large stone sizes, dense composition and the particulars of the pyelocalyceal anatomy.

In our study we have found that stone diameter (p-value 0.001) and the number of stones in pelvis and calices (p-value 0.017) influence stone free rate. Increasing the stone diameter and the number of stones, the probability of a stone free outcome of the RIRS intervention decreases. The calyx anatomy seems not to be particularly relevant. The mean IL is lower (24.5 ± 4.9 vs 25.7 ± 3.8) and the IPA is greater (52.3 ± 24.2 vs 48.0 ± 27.9) in the stone free group than in the non-stone free group, but the difference is statistically not significant. The only thing that seems to have an influence on the success rate is when a stone is present in the lower calix, which has an acute ($< 38.2^0$) infundibulopelvic angle: this negatively influences the stone free rate. In our study we observed that patient with a DJ stent have a high probability of sepsis. Younger and female patients have a higher sepsis rate than older ones and males. The incidence of sepsis in older patients decreases (p-value 0.026). In addition, even in the DJ group women still have a higher probability of having sepsis when compared to men. There are no evident factors that could explain the major incidence in young and female patients.

5 Conclusions

In our study we found that the factors that deteriorate the stone free rate are a higher stone diameter, a larger number of stones and when kidney stones are present in the lower calyx with a sharp infundibular angle. Younger age, being female and having a pre-operative diversion increase the incidence of post-operative sepsis.

References

Turk, C, Knoll, T, Petrik, A and others (2014). *EAU guidelines.*

Wood, SN (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**, 95 – 114.

Modeling Concentration in Data with an Exogenous Order

Andreas Groll[1], Jasmin Abedieh[1], Manuel J. A. Eugster[1]

[1] Department of Statistics, Ludwig-Maximilians-University Munich, Akademiestrasse 1, 80799 Munich, Germany

E-mail for correspondence: andreas.groll@stat.uni-muenchen.de

Abstract: Concentration measures order the statistical units under observation according to their market share. This concept is of great importance and widely applied in practical situations. Over time, certain representatives of classical absolute concentration measures have proven to be useful. However, the formalism of market concentration allows adaption. Here, we present a generalization where an order according to an exogenous variable other than the market share is possible. The resulting generalized concentration index still fulfills the common axioms of (classical) concentration measures and, hence, can still be interpreted like those with the benefit of more precise interpretations and conclusions.

Keywords: Market concentration; Herfindahl index; Rosenbluth index; German Bundesliga; R programming language.

1 Introduction

The concept of market concentration is generally well-established and has become of great importance as it is applied in a variety of fields such as business economics, sociology, sports and many more. Well-established concentration measures like the concentration ratio and the Herfindahl index are computed on the statistical units ordered according to their market share in terms of “specific goods”. From an economics point of view this modus operandi has nice properties; Saving (1970), for example, shows the relation between the Lerner measure of the degree of monopoly and concentration ratios expressed by the market share of the g largest enterprises.

However, to the best of our knowledge, existing concentration concepts do not generalize to a second variable of interest, similar, e.g., to correlation or regression concepts. The central task of this work is to show that the basic concept of measuring concentration can be extended within its classical framework in order

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

to take a second variable into account. This offers room for answering new questions and enables alternative interpretations within the framework of measuring concentration. We generalize the definition of market concentration and allow an order of the enterprises according to a second exogenous variable. Here, "exogenous" simply means "some other (external) variable" and does not concern any information regarding independence or lack of correlation. This exogenous order can be defined by any property of the enterprises—for example, the number of employees, a rating agency's ranking, the geographical position from south to north, or the enterprises' environmental dues. We introduce an appropriate concentration measure. Its main benefit is an additional gain of knowledge as it measures concentration with respect to this exogenous variable. In certain situations, this is desirable and more meaningful than conventional measures. For example, instead of the common analysis of a financial inequality concerning the transaction volume among a group of enterprises, the new measure allows to analyze this financial inequality with respect to e.g., the enterprises' number of employees, rating, market share, etc.

This generalization is a straightforward extension of the classical formalism and we provide the formal concept definition. The rigorous proofs of the axiomatic of (classical) concentration measures, and the implementation and application in the R programming language (R Core Team, 2012) are found in Abedieh et al. (2013). The validity of these axioms is particularly important as it ensures that the new index still can be interpreted in the classical framework of concentration. We hope that this work encourages users/researchers to think about market concentration and concentration measures as a flexible formalism which can be adopted to specific situations.

2 Concentration measures

In this section we review the formalism of classical absolute concentration measures and show common representatives. Given $1, \ldots, n$ statistical units (e.g., enterprises), let X be a specific characteristic of the statistical units (e.g., market share) and $x_1, \ldots, x_n$ positive realizations (observations). We denote the increasing or, respectively, decreasing order of observations by

$$0 \leq x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)} \text{ and } x^{(1)} \geq x^{(2)} \geq \ldots \geq x^{(n)} \geq 0,$$

With $\sum_{i=1}^{n} x_i > 0$, the corresponding ordered relative sums of observations are defined by

$$p_i := x_{(i)} \Big/ \sum_{j=1}^{n} x_j \qquad \text{and} \qquad c_i := x^{(i)} \Big/ \sum_{j=1}^{n} x_j \,, \quad i = 1, \ldots, n.$$

The vectors $\boldsymbol{p}^{\mathrm{T}} = (p_1, \ldots, p_n)$ and $\boldsymbol{c}^{\mathrm{T}} = (c_1, \ldots, c_n)$ represent the corresponding successive sums. Using this formalism, we are able to define common measures of concentration and present three absolute representatives.

Concentration ratio. The concentration ratio is defined as

$$\mathrm{CR}_g := \sum_{i=n-g+1}^{n} p_i = \sum_{i=1}^{g} c_i, \qquad \mathrm{CR}_g \in [0,1]. \tag{1}$$

Concentration ratios show the extend of control of the g largest statistical objects and illustrate the degree of dominance. Based on the concentration ratios, the inequality can be visualized by the concentration curve.

Concentration curve. Based on CR_g the concentration curve is defined. The height of the curve above any point x on the horizontal axis measures the percentage of the statistical unit's (e.g., enterprises) total size accounted for by the largest (with respect to the variable of interest, e.g., the market share) x units. The curve is therefore continuously rising from left to right, but at a continuously diminishing rate (compare Rosenbluth,1955). Hence, it graphically illustrates the inequality among the statistical units in the sense that the concentration is higher the smaller the area above the curve.

Herfindahl index (Hirschman, 1964). It is defined as

$$\mathrm{H} := \sum_{i=1}^{n} p_i^2 = \sum_{i=1}^{n} c_i^2,$$

and results in values $\frac{1}{n} \leq \mathrm{H} \leq 1$. H is an indicator of the amount of competition among the statistical units, i.e., represents the degree of concentration. In the application example in Abedieh et al. (2013), it is then used as an indicator whether there is a monopoly or a significant competition on the transfer spendings of German Bundesliga soccer teams.

Rosenbluth Index (Rosenbluth, 1955, 1957). It is defined as

$$\mathrm{RB} := 1\Big/\Big(2\sum_{i=1}^{n} i\, c_i - 1\Big) = \frac{1}{2A}, \qquad \text{with} \quad A = \sum_{i=1}^{n} i\, c_i - 0.5, \quad 1 \leq 2\,A \leq n,$$

and results in values $\frac{1}{n} \leq \mathrm{RB} \leq 1$. RB denotes the area above the concentration curve. It constitutes an alternative measure to investigate the absolute concentration of a particular group of statistical units based on the kurtosis of the concentration curve.

For both Herfindahl and Rosenbluth index, normalized versions exist:

$$\mathrm{H}^* := (\mathrm{H} - \frac{1}{n}) \Big/ (1 - \frac{1}{n}), \quad \mathrm{RB}^* := (\mathrm{RB} - \frac{1}{n}) \Big/ (1 - \frac{1}{n}), \quad \mathrm{H}^*, \mathrm{RB}^* \in [0, 1].$$

Also the measures' inverse, $n_{\mathrm{H}} = 1/\mathrm{H}$ and $n_{\mathrm{RB}} = 1/\mathrm{RB}$, are of interest as they can be interpreted as the "equivalent number of equal sized units".

With the formalism introduced above, we now can define a measure based on the corresponding concentration ratios for data with exogenous order.

3 Concentration for data with exogenous order

Given data $x_1, \ldots, x_n$, their order based on an exogenous variable is denoted by

$$x_{[1]}, x_{[2]}, \ldots, x_{[n]}.$$

Analogously, we assume $\sum_{i=1}^{n} x_i > 0$ and define the ordered relative sum of statistical units with respect to the exogenous order:

$$q_i := x_{[i]} \Big/ \sum_{j=1}^{n} x_j\,, \qquad i = 1, \ldots, n.$$

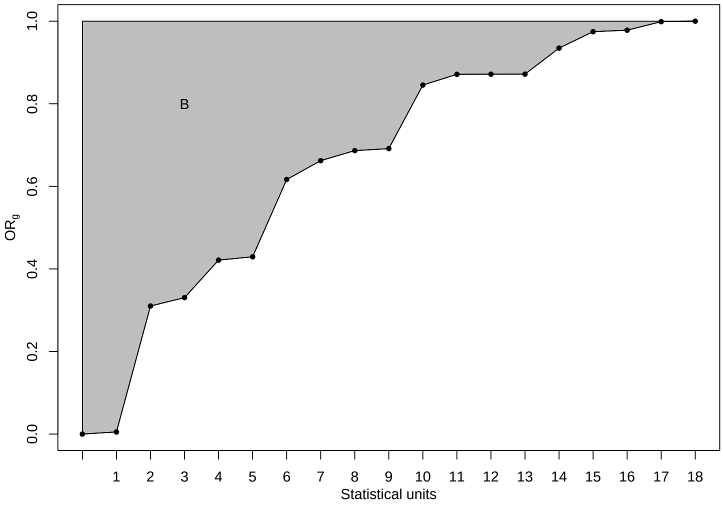


FIGURE 1. Concentration curve for data with an exogenous order.

The vector $\boldsymbol{q}^{\mathrm{T}} = (q_1, \ldots, q_n)$ collects all successive sums.

Concentration ratio. In analogy to the concentration ratio CR_g we define the exogenously ordered concentration ratio; characterizing which part of the sum of objects lies on the group $x_{[1]}, \ldots, x_{[g]}$. We define

$$\mathrm{OR}_g := \sum_{i=1}^{g} q_i, \qquad \mathrm{OR}_g \in [0, 1] .$$

The ratio OR_g allows new ways of interpretation compared to the classical concentration ratio CR_g in (1), as it explains the proportion of the first g statistical units on the overall sum *with respect to the exogenous order.*

Concentration curve. Based on OR_g we define the exogenously ordered concentration curve. In contrast to (classical) concentration curves, ordered relative sums of objects according to the exogenous order form a curve which is still monotone increasing but not necessarily concave. As a consequence the frequency polygon can cross the diagonal from $(0, 0)$ to $(n, 1)$. Figure 1 illustrates a schematic exogenously ordered concentration curve based on transfer spendings of the German Bundesliga (season 2003/04).

Concentration index. We use the schematic exogenously ordered concentration curve in Figure 1 to motivate the definition of an appropriate concentration index. The inequality in data with an exogenous order is illustrated by the area B, which lies above the exogenously ordered concentration curve. The following relation holds: the smaller the surface, the bigger the proportion among the "first few" statistical units. In the extreme case that the whole balance applies on the first statistical unit, one obtains $B_{min} = 0.5$. Note that the uniform distribution does not represent one of the two extreme cases anymore. The inequality among the first statistical units is now minimal, if the whole balance applies to the last statistical unit; in this case we obtain $B_{max} = n - 0.5$. For the uniform distribution we get $q_i = \frac{1}{n} \forall i$ and the exogenous ordered concentration curve is the diagonal with corresponding area $B = n/2$. In general, B is computed by

$$B = \sum_{i=1}^{n} iq_i - 0.5.$$

Based on this area B, we introduce an index which captures the concentration in data with an exogenous order:

$$\mathrm{OI}(n;\boldsymbol{q}) = 1\Big/2B = 1\Big/(2\sum_{i=1}^{n} iq_i - 1)\,, \qquad \mathrm{OI}(n;\boldsymbol{q}) \in [\frac{1}{2n-1}, 1]\,.$$

Now, the uniform distribution with $B = \frac{n}{2}$ results in $\mathrm{OI}(n;\boldsymbol{q}) = \frac{1}{n}$. For the sake of simplicity we set $\mathrm{OI} := \mathrm{OI}(n;\boldsymbol{q})$ wherever dependence on n and $\boldsymbol{q}$ is not crucial. For interpretation, the following statements can be proposed:

$\mathrm{OI} \in (\frac{1}{n}, 1]$	concentration on anterior statistical units
$\mathrm{OI} \in [\frac{1}{2n-1}, 1/n)$	concentration on posterior statistical units
$\mathrm{OI} = 1/n$	no concentration, all statistical units have the same proportion of the sum

In analogy to the classical concentration measures we define a normalized version $\mathrm{OI}^*(n;\boldsymbol{q})$ as well as the measures' inverse. Again, for convenience we use the compact notation $\mathrm{OI}^* := \mathrm{OI}^*(n;\boldsymbol{q})$ wherever the dependence on n and $\boldsymbol{q}$ is not crucial. The normalized version is ($\mathrm{OI}^* \in [0, 1]$):

$$\mathrm{OI}^* := (B - c)\Big/(1 - c) \qquad \text{with} \qquad c = 1\Big/(2n - 1)\,,$$

and the measures' inverse, the "equivalent number of homogenous units", is $n_{\mathrm{OI}} = \frac{1}{\mathrm{OI}}$, with $1 \leq n_{\mathrm{OI}} \leq 2n - 1$. Note that now the interpretation of the measures' inverse n_{OI} is subject to the following restrictions. Here, the extreme cases occur if the whole balance applies to "the first" or "the last" statistical unit (in sense of the exogenous order) resulting in $n_{\mathrm{OI}} \in \{1; 2n - 1\}$. Consequently, the uniform distribution does not represent an extreme case anymore and can be interpreted as a "medium concentration of the first statistical units" with the corresponding equivalent number of homogenous statistical units being exactly equal to the true number of statistical units, $n_{\mathrm{OI}} = n$. Furthermore, it is possible that the equivalent number of homogenous units exceeds the actual number of statistical units, i.e., $n_{\mathrm{OI}} > n$; if this occurs, the balance applies on the last statistical units. In contrast we get $n_{\mathrm{OI}} < n$, if the balance applies on the first statistical units.

Axiomatic of the concentration index OI. Literature discusses and defines a set of characteristics required by absolute concentration measures (see, e.g., Hannah and Kay, 1977; Encaoua and Jacquemin, 1980). The concentration index for exogenous ordered data OI satisfies these axioms—always with respect to the exogenous order and, hence, still can be interpreted in the conventional concentration framework. In Abedieh et al. (2013) we provide the rigorous proofs of the axiomatic and explain them by the help of illustrative examples. Besides, we present the implementation in the R programming language (R Core Team, 2012) together with an application to the transfer spendings of German Bundesliga soccer teams.

References

Abedieh, J., Groll, A., and Eugster, M.J.A. (2013). Measuring Concentration in Data with an Exogenous Order. *Technical Report*, **140**, Ludwig-Maximilians-University.

Encaoua, D. and Jacquemin, A. (1980). Degree of monopoly, indices of concentration and threat of entry. *International Economic Review*, **21(1)**, 87 – 105.

Hannah, L. and Kay, J. A. (1977). *Concentration in Modern Industry: Theory, Measurement and the U.K. Experience.* Macmillan.

Hirschman, A.O. (1964). The paternity of an index. *American Economic Review*, **54**, 761 – 762.

R Core Team (2012). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria, ISBN 3-900051-07-0, `http://www.R-project.org/`.

Rosenbluth, G. (1955). Measures of concentration. In: *National Bureau of Economic Research (Ed.), Business Concentration and Price Policy.* Princeton University Press.

Rosenbluth, G. (1957). *Concentration in Canadian Manufacturing Industries.* Princeton University Press.

Saving, T. S. (1970). Concentration ratios and the degree of monopoly. *International Economic Review*, **11(1)**, 139 – 146.

Impact of misspecified random effects distribution on the variance component test in GLMM

Freddy Hernández[1], Olga Usuga[2], Jorge I. Vélez[3,4]

[1] Universidad Nacional de Colombia, Medellín, Colombia.
[2] Universidad de Antioquia, Medellín, Colombia.
[3] Universidad del Norte, Barranquilla, Colombia.
[4] Arcos-Burgos Group, Department of Genome Sciences, John Curtin School of Medical Research, Australian National University, Canberra, Australia.

E-mail for correspondence: `fhernanb@unal.edu.co`

Abstract: Clustered data are, nowadays, available in multiple disciplines. Statistical modeling of these data is usually performed using Generalized Linear Mixed Models (GLMMs), proposed by Breslow & Clayton (1993), to account for the hierarchical structure in the data. One problem of interest is to determine whether, in the fitted model, the variance component (σ^2) of the random effects is statistically zero (that is, the random effect does not explain much of the variance and should be excluded from the model). In the literature, this feature is assessed using a variance component test. Here we present the results of a statistical simulation study assessing the performance of two variance component tests in the context of GLMMs, when a binary and Gamma response variables are considered. In particular, the likelihood ratio (LR) and the permutation tests are evaluated, and the impact of misspecifying the true distribution of the random effects is measured as a function of the number of individuals per cluster and the true value of σ^2. We found that, for both response variables, the LR and permutation tests are affected for the number of individuals per cluster and the prespecified value of σ^2, but have a similar behavior regardless of the true distribution of the random effects.

Keywords: Mixed models; variance components; statistical computing.

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

1 Introduction

Clustered data are commonly collected in studies on medical, social, and behavioral sciences. This type of data arise when there is a hierarchical or clustered structure in the data such as individuals nested in institutions or organizations (i.e., students in schools, employees in firms, or patients in hospitals). Mixed models, hierarchical models or multilevel regression models provide an attractive framework to accommodate the overdispersion and dependence of this type of data (Zhu & Zhang, 2006). Many of the models used in these fields fall under the frame of Generalized Linear Mixed Models (GLMMs). A fundamental question in this models is about the heterogeneity among clusters, which is equivalent to test whether σ^2, the variance component associated to the random effects, is statistically zero. This test is known as a *variance component test*, and can be approached using the likelihood ratio (LR) and permutation tests (LRs), and have important implications in statistical modeling of clustered data. Here we use a statistical simulation approach to determine, which of the aforementioned tests is more appropriate to determine whether σ^2 is effectively zero. To do this, different statistical distributions and several prespecified values of σ^2 are utilized to generate the vector of random effects, followed by the underlying data to be modeled.

2 GLMMs

GLMMs, proposed by Breslow & Clayton (1993), have been extensively in many applications where clustered and/or longitudinal data are available. Let y_{ij} the jth response variable within the ith cluster ($i = 1, 2, \ldots, m; j = 1, 2, \ldots, n_i$). In a GLMMs with random intercept, it is assumed that, conditional to the random effects b_i, the outcome y_{ij} is independent with the following structure:

$$\begin{aligned} y_{ij} \mid b_i &\sim \text{independent in } F_y, \\ g(\mu_{ij}) &= \boldsymbol{X}_{ij}\boldsymbol{\beta} + b_i, \\ b_i &\overset{ind}{\sim} N(0, \sigma^2), \end{aligned} \tag{1}$$

where F_y belongs to the exponential family, $g(\cdot)$ is a known link function, $\boldsymbol{X}_{ij}$ is the vector of covariates for the jth observation in the i cluster, and $\boldsymbol{\beta}$ is the parameter vector for the fixed effects.

3 Variance component test

Among the tests available in the literature for testing

$$H_0 : \sigma^2 = 0 \text{ vs. } H_A : \sigma^2 > 0, \tag{2}$$

where σ^2 is the variance of random intercept, the LR test is the most commonly used test because of its theoretical properties and straightforward construction. In a GLMM with random intercept, we are interested in testing H_0 in (2) using a type I error probability α. The statistic for the LR test is calculated as $T =$

$-2\log(L_0/L_A)$ where L_0 and L_A are the likelihood model under H_0 and H_A, respectively. Under H_0, the asymptotic null distribution of T is a GLMM with random intercept is a 50:50 mixture between a χ_0^2 and χ_1^2 distributions (Zhang & Lin, 2008). The permutation test (PT) is a modification of LR test proposed by Fitzmaurice & Lipsitz (2007). The idea behind the PT is that, when H_0 in (2) is true, the heterogeneity between clusters is non-existing. This result implies that we could mix the clusters without changing the decision. In what follows we present the implementation of the PR given by Fitzmaurice & Lipsitz (2007). First, calculate the LR test statistic in the original sample and denote it by T_{obs}. Second, permute the cluster indexes while holding fixed the number of units within a cluster, n_i, and calculate the LR test statistic T. Third, repeat step 2, M times, to obtain $T_1, T_2, \ldots, M$, where T_m is the LR statistic for the mth permutation of the original clusters ($b = 1, 2, \ldots, M$). Fourth, determine the PT p-value as the proportion of permutation samples where $T_i \geq T_{obs}$ for $i = 1, 2, \ldots, M$.

4 Simulation study

In order to compare the two variance component tests, we consider a GLMM with random intercept and a response variable following a binary and Gamma distributions. In both cases, the random intercept was sampled from four true statistical distributions (Normal, Log-Normal, Exponential and Uniform). For the fitting procedure, normality of the random intercept was assumed. The model considered in this case can be summarised as follows:

$$\text{logit}\{P(y_{ij} = 1 \mid b_i)\} = \beta_0 + \beta_{between} x_1 + \beta_{within} x_2 + b_i, \tag{3}$$

where $i = 1, 2, \ldots, m$ represents the cluster and $j = 1, 2, \ldots, n_i$ represents the number of observations per cluster. The between-cluster covariate $x_1 \sim$ Poisson($\lambda = 2$) and x_2 is a within-cluster covariate following a $U(0, 1)$ distribution. The true model parameters are $\beta_0 = -2.5$, $\beta_{between} = 2$ and $\beta_{within} = 1$. Figure 1 displays the main results for the binary GLMM outlined in §3. Overall, the average rejection rate (ARR) of H_0 in (2) of the LR test and the PT are affected for the number of individuals per cluster (parameter n_i in our simulation approach) and the pre-specified value of σ^2, but have a similar behaviour regardless of the true distribution of the random effects. To directly compare the ARRs between the LR test and PT, the ratio $\gamma = \text{ARR}_{\text{LR}}/\text{ARR}_{\text{PT}}$ was calculated. Here, values of $\gamma > 1$ indicate that the LR test outperforms the PT; values of $\gamma < 1$ indicate that the LR test outperforms the PT, and $\gamma = 1$ indicate that the LR test and PT produce equivalent rejection rates. When the number of individuals per cluster is small (that is, $n_i < 5$), the PT has a higher ARR than the LR test (see third column in Figure 1). Our results show that the LR test and PT are not significantly affected by any of the aforementioned parameters, but that the PT outperforms the LR test. This result has important implications when modelling clustered and/or longitudinal data.

References

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

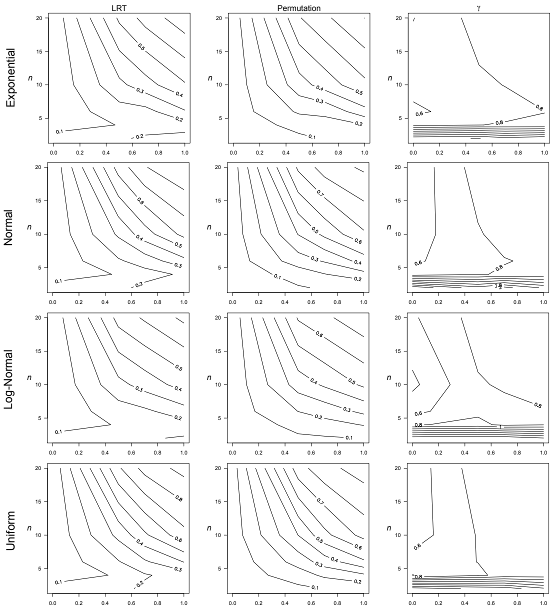


FIGURE 1. ARR of $H_0 : \sigma^2 = 0$ in (2) in the binary GLMM as a function of n_i and σ^2 for each distribution of the random effects. The third column corresponds to γ.

9 – 25.

Fitzmaurice, G.M. and Lipsitz, S.R. (2007). A Note on Permutation Tests for Variance Components in Multilevel Generalized Linear Mixed Models. *Biometrics*, **63**, 942 – 946.

Zhang, D. and Lin, X. (2008). *Variance Component Testing in Generalized Linear Mixed Models for Longitudinal/Clustered Data and other Related Topics* in Dubson, D.B. (2008) Random Effect and Latent Variable Model Selection. New York: Springer.

Modeling data with truncated and inflated Poisson distribution

Ting Hsiang Lin[1], Min-Hsiao Tsai[1]

[1] National Taipei University, Taiwan, Republic of China

E-mail for correspondence: `tinghlin@mail.ntpu.edu.tw`

Abstract: Zero inflated Poisson regression is a commonly used model to analyze data with excessive zeros. Although many models have been developed to fit zero-inflated data, many of them strongly depend on special features of the individual data. For example, there is a need for new models when dealing with truncated and inflated data. In this paper, we proposed a new model with flexibility to model inflations and truncations simultaneously, and the model is a mixture of multinomial logistic and truncated Poisson regression, in which the multinomial logistic component models the occurrence of excessive counts. The truncated Poisson regression models the counts that are assumed to follow a truncated Poisson distribution. The performance of our proposed model is evaluated through simulation studies, and our model has smallest mean absolute error and best model fit. In the empirical example, the data is truncated with inflated zeros and fourteen and the result showed that our model exhibited a better fit than other competing models.

Keywords: zero-inflated data; truncated data; Poisson regression.

1 Introduction

For zero-inflated data, zero-inflated Poisson model (ZIP; Lambert, 1992) and its variants have become a popular tool for analyzing count data with excessive zeros. The problem with ZIP is it can only model a single inflated value and the inflated value has to be zero. Model with other inflated value other than zero have been proposed by Famoye and Singh (2003), Bae and Famoye (2005) and they considered data with a massive point K. The ZIP has been extended to a ZKIP model by Lin and Tsai (2014) that models data with masses of zero and K concurrently. The model is a mixture of multinomial logistic and Poisson regression, in which the multinomial logistic component models the occurrence of excessive counts, including zeros, K (where K is a positive integer) and all other values.

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

Truncated data occurs when the subjects are observed within a certain time window. A subject whose event is not in this time interval will be not observed and no information on this subject is available. Because the data is only available within the observational window, the inference for truncated data is restricted to conditional estimation. When the end point of the time window is defined, the data is right truncated. For example, “Exactly, how many times have you been laid off?” In this type of truncation, any subjects who experience the event of interest after the truncation time are not observed and some justifications have to be made to ensure the correct predictions beyond the time frame. In this study, we extend the existing models through the flexibility of modeling inflations and truncations of the data with excessive counts other than zeros. The model we propose is a mixture of multinomial logistic and truncated Poisson regression, while the multinomial logistic component models the occurrence of excessive counts, including zero and K. The truncated Poisson regression component models the counts that are assumed to follow a truncated Poisson distribution.

2 Main Result

In this study, we proposed new models to fit data with truncated and inflated values. We extended ZIP model for truncated Poisson distribution and proposed an inflated truncated Poisson regression model. The model has two variations under different conditions. The first derivation is zero inflated truncated Poisson regression (ZITP) with one inflated point at zero and truncated at K. The second derivation is zero-K inflated truncated Poisson (ZKITP) with two inflated and truncated points: zeros and K. These models can be considered mixture models of two parts. The first part of the models is to fit whether the inflated values occur or not, and it is fitted by a binary logistic model for ZITP, and a multinomial logistic model for ZKITP. The second part of the models handles the non-inflated counts, which is fitted by a truncated Poisson regression model. We use data from the Behavioral Risk Factor Surveillance System (BRFSS) to compare the fit of the TP (truncated Poisson) model to those of ZITP and ZKITP. The variable used to demonstrate the models was constructed from the question: “Over the last 2 weeks, how many days have you had trouble falling asleep or staying asleep or sleeping too much?” There were 2171 subjects with valid responses, and 1322 that reported zero (60.89%), 293 reported 14 (13.50%), and the rest reported between 1 and 13 days. In our case, the data were fitted to the TP, ZITP and ZKITP models, with the age of the respondents as a predictor. The results of the model fit and their MAE, MSE are presented in Table 1. Since the models are nested, a likelihood ratio test can be performed for model comparison. Table 2 shows that the log-likelihood comparison of TP vs ZITP is 8660.442, and ZITP vs ZKITP is 2872.215, and the degrees of freedom differ all by 2. The difference of twice the negative log-likelihood between 2 models follows a chi-square distribution, and the results showed ZITP outperformed TP, and ZKITP significantly improved ZITP even further. The mean absolute error (MAE) is smallest for ZKITP, suggesting that ZKITP is better at making predictions than the other two models. Figure 1 displays the predicted values by number of days had trouble falling asleep, and the results show that the predicted values of ZKITP are closer to the observed frequencies compared to the ZITP and TP for most situations, particularly for

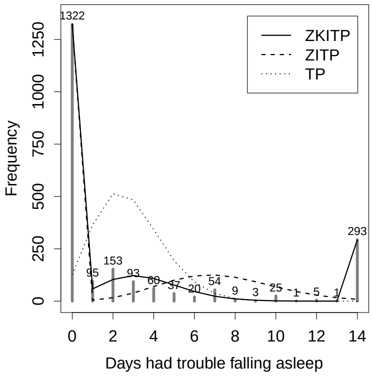


FIGURE 1. The observed and predicted values by different models.

the two inflated values 0 and 14. The performances of ZITP is better than TP at fitting zero, but not for 14, and TP fit poorly for both values.

TABLE 1. Performance of the three models.

Index	ZKITP	ZITP	TP
AIC	6524.752	9392.968	18049.410
BIC	6558.850	9415.699	18060.775
MAE	19.506	75.395	204.528
MSE	717.632	10233.750	131624.000

TABLE 2. Model comparison of the three models.

Model Comparison	LR-Test Statistic	DF	P-Value
TP vs ZITP	8660.442	2	< 0.0001
ZITP vs ZKITP	2872.215	2	< 0.0001

References

Bae, S., Famoye, F., Wulu, J.T., Bartolucci, A.A. and Singh, K.P. (2005). A rich family of generalized Poisson regression models with applications. *Mathematics and Computers in Simulation*, **69**, 4 – 11.

Famoye, F. and Singh, K.P. (2003). On Inflated generalized Poisson regression models. *Advances and Applications in Statistics*, **3**, 145 – 158.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1 – 14.

Lin, T.H. and Tsai, M.H. (2014). Modeling health survey data with excessive zero and K responses. *Statistics in Medicine*, **32**, 1572 – 1583.

Saffari, S.E., Adnan, R. and Greene, W. (2011). Handling of over-dispersion of count data via truncation using Poisson regression model. *Journal of Computer Science and Computational Mathematics*, **1**, 1 – 4.

Wang, H. and Heitjan, D.F. (2008). Modeling heaping in self-reported cigarette counts. *Statistics in Medicine*, **27**, 3789 – 3804.

Welsh, A.H., Cunningham, R.B., Donnelly, C.F. and Lindenmayer, D.B. (1996). Modelling the abundance of rare species: Statistical models for counts with extra zeros. *Ecological Modelling*, **88**, 297 – 308.

Xie, T. and Aickin, M. (1997). A truncated poisson regression model with applications to occurrence of adenomatous polyps. *Statistics in Medicine*, **16**, 1845 – 1857.

Zhou, X.H. and Tu, W.Z. (1999). Comparison of several independent population means when their samples contain log-normal and possibly zero observations. *Biometrics*, **55**, 645 – 651.

Bayesian joint spatial models for intimate partner violence and child maltreatment

Antonio López-Quílez[1], Miriam Marco[2], Enrique Gracia[2], Marisol Lila[2]

[1] Dept. of Statistics and Operational Research, University of Valencia, Spain
[2] Dept. of Social Psychology, University of Valencia, Spain

E-mail for correspondence: antonio.lopez@uv.es

Abstract: In this study we propose Bayesian hierarchical models to analyse the geographical distribution of intimate partner violence and child maltreatment. We take into account contextual factors that may explain these distributions and also spatially structured and unstructured latent variables. The relationship between both types of violence is also explored with shared component models.

Keywords: Ecological regression; Family violence; Multivariate spatial models.

1 Introduction

Disease mapping and spatial regression with count data are widely used in epidemiology (Lawson et al., 1999). An increasing number of approaches have been recently proposed for the analysis of spatially aggregated health count data with refinements and generalizations. Those complex models open new possibilities in other knowledge areas as for example the social sciences. Research from a spatial perspective has shown social problems exhibit geographic variation. A growing number of studies are using spatial Bayesian methods in the field of crime and social problems with promising results (Law et al., 2014; Gracia et al., 2014).

2 Modelling family violence

We used as neighbourhood units the 552 census block groups in the city of Valencia to study the geographical distribution of intimate partner violence and child maltreatment. A Bayesian random-effects modelling approach was used to analyse the influence of neighbourhood-level characteristics on small-area variations in both aspects of family violence.

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

Data for each response variable were counts for 552 census block groups; therefore we assume a Poisson regression (Gracia et al., 2015):

$$y_i|\eta_i \sim Po(E_i \exp(\eta_i)),$$

where the log-relative risk η_i is explained by the covariates X, a spatially structured term S and an unstructured term U, in the following form

$$\eta_i = \mu + X_i\beta + S_i + U_i.$$

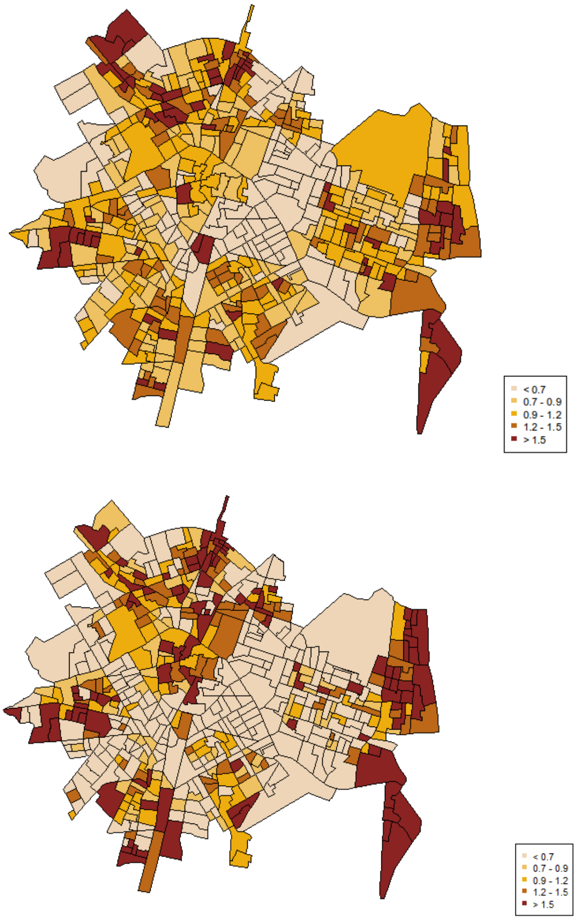


FIGURE 1. Maps of relative risks of intimate partner violence (up) and child maltreatment (down) in the city of Valencia.

By estimating spatially structured and unstructured random effects, we aimed to assess separately the influences of spatial dependency and independent heterogeneity. Figure 1 shows the maps of intimate partner violence and child maltreatment risks respectively. They present some similarities and also differences.

Figure 2 shows the posterior mean of the spatial component of intimate partner violence risk and child maltreatment. The geographical patterns are clearly different, but exhibit a certain complementarity. These results suggest the existence of a relationship between the two processes.

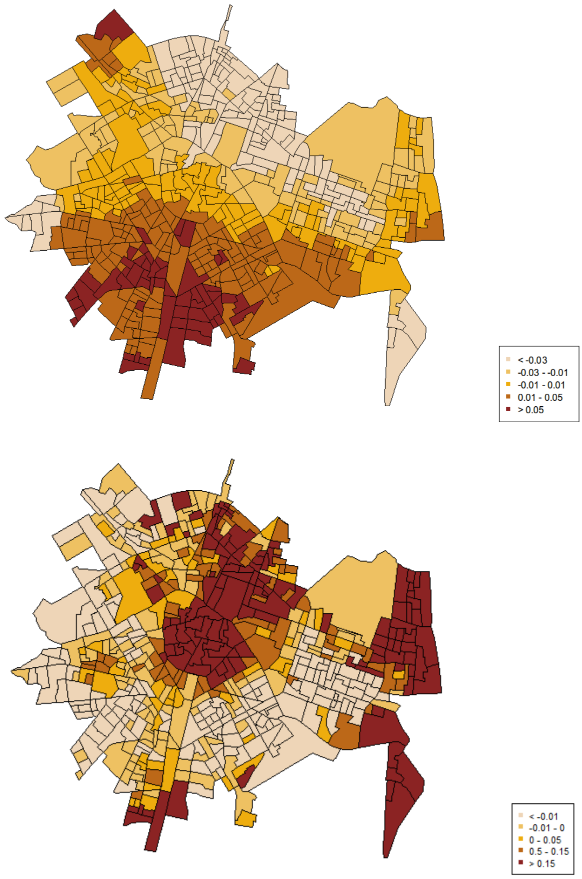


FIGURE 2. Posterior mean of the spatial component of intimate partner violence (up) and child maltreatment (down).

3 Multivariate spatial modelling

A joint spatial modelling of intimate partner violence risk and child maltreatment is explored in order to assess the common structure and the particularities of each kind of family violence.

Knorr-Held and Best (2001) analysed health data using a shared component model. This model is similar in spirit to conventional factor analysis, and parti-

tions the geographical variation of two response variables into a common (shared) component and two response-specific (residual) terms.
Based on social disorganization theory combined with the available data, several covariates at the neighbourhood level are used in the modelling. Different approaches for incorporating the covariates are studied and compared.

Acknowledgments: This research was supported by the Spanish Ministerio de Economía y Competitividad (PSI2014-54561-P and MTM2013-42323-P). Miriam Marco was supported by the FPU program of the Spanish Ministerio de Educación, Cultura y Deporte (FPU2013/00164).

References

Gracia, E., López-Quílez, A., Marco, M., Lladosa, S., and Lila, M. (2014). Exploring neighborhood influences on small-area variations in intimate partner violence risk: A bayesian random-effects modeling approach. *International Journal of Environmental Research and Public Health*, **11**, 866 – 882.

Gracia, E., López-Quílez, A., Marco, M., Lladosa, S., and Lila, M. (2015). The spatial epidemiology of intimate partner violence: Do neighborhoods matter? *American Journal of Epidemiology*, **182**, 58 – 66.

Knorr-Held, L. and Best, N.G. (2001). A shared component model for joint and selective clustering of two diseases. *Journal of the Royal Statistical Society, Series A*, **164**, 73 – 85.

Law, J., Quick, M., and Chan, P. (2014). Bayesian spatio-temporal modeling for analysing local patterns of crime over time at the small-area level. *Journal of Quantitative Criminology*, **30**, 57 – 78.

Lawson, A.B., Biggeri, A.B., Böhning, D., Lesaffre, E., Viel, J.F. and Bertollini, R. (eds.) (1999). *Disease Mapping and Risk Assessment for Public Health.* New York: John Wiley and Sons.

Variations of compound models

Wan Jing Low[1], Paul Wilson[1], Mike Thelwall[1]

[1] School of Mathematics and Computer Science, University of Wolverhampton, United Kingdom

E-mail for correspondence: W.J.Low@wlv.ac.uk

Abstract: We extend the analysis of discrete data by comparing two new variations of compound models with the traditional model on a standard data set. The results suggest that the new models have similar fits to the traditional model, especially in terms of log likelihood.

Keywords: Compound variant; Poisson; Negative binomial

1 Introduction

Compound distributions (also known as stopped sum distributions) are discrete distributions that are used in applications such as the branching process in ecology (Neyman, 1939) and for risk assessment (Ager and Norman, 1964). Using an ecological example, say there are N parent insects in a field, which follow a discrete distribution, and these parents then independently give rise to offspring (X_i), which follow another distribution. The total number of offspring, S_N, follow a distribution of the form:

$$S_N = X_1 + X_2 + \cdots + X_N$$

where $i = 1, 2, 3, ..., N$, and N is a random draw. The use of terms varies slightly between authors, as this has been referred to as compound, mixture and stopped sums in different cases (Johnson et al. 2005). Here we use the term ‘compound’ to denote such a distribution. A common compound distribution is the Neyman Type A (compound Poisson-Poisson), in which the parent and offspring generations follow Poisson distributions with parameters λ and ϕ respectively, with probability mass function (p.m.f.):

$$P(X = x) = \frac{\mathrm{e}^{-\lambda}\phi^x}{x!}\sum_{j=0}^{\infty}\frac{\left(\lambda \mathrm{e}^{-\phi}\right)^j j^x}{j!}$$

We propose two variants that model the sum of parent and offspring generations.

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

2 Compound model variants

There is already a compound variant which conforms to the zero restriction assumption in the 'traditional' compound model, where a zero in the first generation will automatically result in a zero in the second generation. Unlike the 'traditional' compound model, this variant models the sum of parent and offspring generations, rather than just the offspring. This variant, known as SVA, was previously introduced by Low et al. (2016). For example, the p.m.f. for SVA Poisson-Poisson, (where λ and ϕ are the Poisson parameters for the two generations) is:

$$P(X = x) = \begin{cases} \mathrm{e}^{-\lambda} & x = 0 \\ \displaystyle\sum_{j=1}^{x} \frac{\mathrm{e}^{-\lambda}\lambda^{j}}{j!} \frac{\mathrm{e}^{-\phi}\phi^{x-j}}{(x-j)!} & x \neq 0 \end{cases}$$

The negative binomial (NB) model is commonly used to address overdispersion, which is common in many discrete data sets. We also consider here three other cases (in all cases below μ and α are the mean and dispersion parameters of the NB distribution, with $p = \alpha/(\mu + \alpha)$ and $q = 1 - p$):

1. SVA Poisson-NB,

$$P(X = x) = \begin{cases} \mathrm{e}^{-\lambda} & x = 0 \\ \displaystyle\sum_{j=1}^{x} \frac{\mathrm{e}^{-\lambda}\lambda^{j}}{j!} \binom{x-j+\alpha-1}{\alpha-1} p^{\alpha} q^{x-j} & x \neq 0 \end{cases}$$

2. SVA NB-Poisson,

$$P(X = x) = \begin{cases} p^{\alpha} & x = 0 \\ \displaystyle\sum_{j=1}^{x} \binom{j+\alpha-1}{\alpha-1} p^{\alpha} q^{j} \frac{\mathrm{e}^{-\lambda}\lambda^{x-j}}{(x-j)!} & x \neq 0 \end{cases}$$

3. SVA NB-NB,

$$P(X = x) = \begin{cases} p^{\alpha} & x = 0 \\ \displaystyle\sum_{j=1}^{x} \binom{j+\alpha-1}{\alpha-1} p^{\alpha} q^{j} \binom{x-j+\theta-1}{\theta-1} r^{\theta} s^{x-j} & x \neq 0 \end{cases}$$

In SVA NB-NB, the second NB distribution has parameters μ_2 and θ, where $r = \theta/(\mu_2 + \theta)$ and $s = 1 - r$

There is a second variant in which it is possible to obtain a non-zero second generation even if a zero is obtained in the first, because in some applications this situation is possible. The second variant is called SVB. For example, the SVB Poisson-NB has p.m.f.:

$$P(X = x) = \sum_{j=0}^{x} \frac{\mathrm{e}^{-\lambda}\lambda^{j}}{j!} \binom{x-j+\alpha-1}{\alpha-1} p^{\alpha} q^{x-j}$$

We also consider SVB NB-NB, which has p.m.f.:

$$P(X = x) = \sum_{j=0}^{x} \binom{j+\alpha-1}{\alpha-1} p^{\alpha} q^{j} \binom{x-j+\theta-1}{\theta-1} r^{\theta} s^{x-j}$$

Although SVB Poisson-NB and SVB NB-Poisson are the same, in general the order of the distributions within the compound variants does matter.

3 Application and methods

Previous research (Low et al., 2016) shows that when these variant models are fitted to data sets consisting of citation counts for sets of articles from the same discipline and year, in many cases the AIC obtained is lower. Here, we further investigate these proposed models using biodosimetry data previously collected by Romm et al. (2013) and also analysed by Oliveira et al. (2016). The data set contains the frequency of automatically detected dicentric chromosomes, which were exposed to eight uniform doses of Cobalt-60 gamma rays. Similar to the 'traditional' compound model, our proposed models also account for the overdispersion, but have greater flexibility, thus could provide a suitable fit for this data. In all cases, a log-link and the quadratic model:

$$\textit{Mean number of dicentric chromosomes} \sim \textit{dose} + \textit{dose}^2$$

are used, allowing us to obtain results that are comparable with those of Oliveira et al. (2016). All models are fitted using code written by the first author in R (R Development Core Team, 2014).

4 Results and conclusion

TABLE 1. 'Traditional' compound models fitted to biodosimetry data.

Models	Parameters	Log likelihood	AIC	BIC
Neyman type A	6	−3738.21	7488	7534
Compound Poisson-NB	9	−3734.25	7486	7555
Compound NB-Poisson	9	−3739.43	7497	7566
Compound NB-NB	12	−3739.57	7493	7585

Based on our initial results, the 'traditional' compound Poisson-NB gave the lowest AIC, whilst Neyman type A produced the lowest BIC. However, the SVA Poisson-Poisson gave the third lowest BIC (see Table 2). Overall our proposed models have similar fits to the previously used model, especially in terms of log likelihoods. However, the extra parameters in the variant models are penalised, resulting in larger AICs. Diagnostic plots of distributions of residuals show that the variant models and the 'traditional' models fit equally well. Therefore, we recommend testing these models in the analysis of data with similar properties in the future, in case they fit better for such data sets.

TABLE 2. SVA and SVB models fitted to biodosimetry data.

Models	Parameters	Log likelihood	AIC	BIC
SVA Poisson-Poisson	6	−3749.36	7511	7557
SVA Poisson-NB	9	−3749.36	7517	7586
SVA NB-Poisson	9	−3741.47	7501	7570
SVA NB-NB	12	−3744.89	7514	7606
SVB Poisson-NB	9	−3749.36	7517	7586
SVB NB-NB	12	−3747.72	7519	7611

References

Ager, J. E. and Norman, L. G. (1964). The causation of bus driver accidents. *Occupational and Environmental Medicine*, **21**, 248 – 248.

Johnson, N.L., Kemp, A. W. and Kotz, S. (2005). *Univariate Discrete Distribution.* 3rd ed. New Jersey: Wiley & Sons.

Low, W.J., Wilson, P. and Thelwall, M. (2016). Stopped sum models and proposed variants for citation data. *Scientometrics*, **107**, 369 – 384.

Neyman, J (1939). On a new class of "contagious" distributions, applicable in entomology and bacteriology. *The Annals of Mathematical Statistics*, **10**, 35 – 57.

Oliveira, M., Einbeck, J., Higueras, M., Ainsbury, E., Puig, P. and Rothk- amm, K (2016). Zero-inflated regression models for radiation-induced chromosome aberration data: A comparative study. *Biometrical Journal*, **58**, 259 - 279.

R Development Core Team (2014). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, AT.

Romm, H., Ainsbury, E., Barnard, S., Barrios, L. and 14 others. (2013). Automatic scoring of dicentric chromosomes as a tool in large scale radiation accidents. *Mutation Research*, **756**, 174 – 183.

Zero inflated hyper-Poisson regression model

Ana M. Martínez-Rodríguez[1], Antonio Conde-Sánchez[1], Antonio J. Sáez-Castillo[1]

[1] Department of Statistics and Operational Research, Universidad de Jaén, Spain

E-mail for correspondence: ammartin@ujaen.es

Abstract: The zero inflated hyper-Poisson regression model permits count data to be analysed with covariates that determine different levels of dispersion and that present structural zeros due to the existence of a non-users group. An application of the model to fit frequency of doctor visits in relation to several covariates confirms the presence of structural zeros at the same time as it detects under-dispersion in some of the levels determined by the covariates.

Keywords: Count data; Hyper-Poisson; Under-dispersion; Zero inflation.

1 Introduction

Over-dispersion phenomenon in probability distributions related to count data refers to the presence of variability in data higher than that corresponding to the Poisson distribution. A common over-dispersion source is the existence of an extra amount of zeros in relation to the number of zeros that a Poisson distribution may present. That extra amount of zeros may appear when the dataset is divided in two populations: in the first, the counting event is impossible (so it is called the *non-users* group) and all the variable values are *structural zeros*; in the second (the *potential users* group), the event may affect the individuals, so a zero value may appear but also higher values. In order to analyse that kind of datasets, zero inflated models propose a mixture of two processes: on the one hand, it considers that a binary process generates the structural zeros; on the other hand, when a datum is not an structural zero, a counting process generates it. A logistic model or a censured counting process may be considered for the binary process, whereas the counting process is commonly modelled by a Poisson or a negative binomial distribution, giving way to the Zero Inflated Poisson (ZIP) or the Zero Inflated Negative Binomial (ZINB) models, respectively.
Although over-dispersion is the most common situation in count data, under-dispersion, that appears when the variance is below the mean, is also possible.

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

That under-dispersion phenomenon can be explained by the existence of a negative contagion effect in the counting process.

Modelling in an under-dispersion context gets more complex if the dataset presents structural zeros: in that case, ZIP or ZINB models cannot work adequately, because they would only be able to reflect equi-dispersion or over-dispersion in the potential users group, and a model that can manage under-dispersion is necessary. Moreover, Sellers and Shmueli (2013) have shown that datasets may contain mixtures of populations, some of them being over-dispersed and others under-dispersed.

That possibility of the misspecification is more evident when these data are zero inflated, because considering data from both groups jointly provides over-dispersion, hiding the fact that some cases in the potential users group, or even all of them, may be under-dispersed.

Therefore, our goal in this work is to study the capability of a zero inflated model where the probability distribution for the counting process of the potential users group is a hyper-Poisson distribution, to model a dataset detecting the presence of a non-users group and, at the same time, being able to distinguish over- and under-dispersion in the group of potential users. The model is described in section 2. Section 3 includes an application to real data.

2 The Zero Inflated hyper-Poisson regression model

Let y_i being the value of the response variable of the $i-$th individual of the sample, a zero inflated model for it may be defined by

$$Y_i = p_i^* Y_i^*,$$

where p_i^* is a binary realization of a Bernoulli variable with probability p_i, which represents the probability of the $i-$th individual to be a structural zero, and Y_i^* is the count variable for non-structural zero individuals of the sample.

The probability mass function (p.m.f.) is then

$$P\left[Y_i = y_i\right] = \begin{cases} p_i + \left(1-p_i\right) P\left[Y_i^* = 0\right] & \text{if } y_i = 0 \\ \left(1-p_i\right) P\left[Y_i^* = y_i\right] & \text{if } y_i > 0 \end{cases}.$$

We define the Zero Inflated hyper-Poisson model (from now on, ZIhP) as the zero inflated model where the distribution of Y_i^* is a hyper-Poisson distribution, whose p.m.f. is

$$P\left[Y_i^* = y_i\right] = \frac{1}{{}_1F_1\left(1;\gamma_i;\lambda_i\right)} \frac{\lambda_i^{y_i}}{\left(\gamma_i\right)_{y_i}},\ y_i = 0, 1, ...,$$

being $(a)_r = a\left(a+1\right)...\left(a+r-1\right),\ a > 0,\ r$ a positive integer, and ${}_1F_1\left(a;b;c\right)$ the confluent function. γ_i is a dispersion parameter which determines that the distribution is over-dispersed if $\gamma_i > 1$, under-dispersed if $\gamma_i < 1$ and if $\gamma_i = 1$ it matches with the Poisson. λ_i is interpreted as a location parameter.

The model may include covariates both in p_i, and Y_i^*. Let we denote $\mathbf{x_i^{\mathrm{T}}} = (1, x_{i1}, x_{i2}, ..., x_{ik})$ the observed covariates that affect Y_i^* and $\mathbf{z_i^{\mathrm{T}}} = (1, z_{i1}, z_{i2}, ..., z_{il})$ those that affect p_i. Then, we consider

$$p_i = \frac{1}{1+\exp\left(-\mathbf{z_i^{\mathrm{T}}}\boldsymbol{\nu}\right)}.$$

And we choose the formulation of the hyper-Poisson regression model (Sáez-Castillo and Conde-Sánchez, 2013) for Y_i^* modelling the mean as

$$\mu_i = E\left[Y_i\right] = \exp\left(\mathbf{x_i}^{\mathrm{T}}\boldsymbol{\beta}\right)$$

and, optionally, from the dispersion parameter as

$$\gamma_i = \exp\left(\mathbf{x_i}^{\mathrm{T}}\boldsymbol{\delta}\right). \tag{1}$$

Since it is known that

$$\mu_i = \lambda_i - (\gamma_i - 1)\frac{{}_1F_1\left(1;\gamma_i;\lambda_i\right) - 1}{{}_1F_1\left(1;\gamma_i;\lambda_i\right)}, \tag{2}$$

it is possible to obtain the location parameter λ_i for each case solving (2) once we have determined μ_i and γ_i. As we have mentioned, the model determines an under-dispersed, over-dispersed or equi-dispersed distribution for Y_i^* depending on γ_i value.

So it is possible to detect different variance-means ratios within the cases, depending on covariates and once the effect of structural zeros has been isolated.

We estimate the model coefficients by maximizing the log-likelihood function (see Sáez-Castillo and Conde-Sánchez, 2016 for details).

3 Application to real data

We illustrate the use of the ZIhP with real data by considering a dataset from the Australian Health Survey 1977-1978 used by Staub and Winkelmann (2013). The dependent variable is the number of consultations with a doctor or specialist in the 2-week period before the interview. Regressors include demographics (sex and age), income, various measures of health status (number of reduced activity days, general health questionnaire score, recent illness, chronic condition 1 and chronic condition 2) and four types of health insurance coverage (Levyplus, Freepor and Freerepat, with Levy the omitted category).

Table 1 shows the obtained results for the ZIhP model. Only significant covariates for γ have been considered. The ZINB model has an AIC of 6269.187 whereas de ZIhP has an AIC of 6228.763, so the proposed model produces a better fit. In addition, it can be checked by replacing the estimates in (1) that on most occasions $\gamma_i < 1$, thus under-dispersion is present. Therefore, in this dataset ZINB model cannot work adequately because it only is able to reflect over-dispersion.

References

Sáez-Castillo, A.J. and Conde-Sánchez, A. (2013). A hyper-Poisson regression model for overdispersed and underdispersed count data. *Computational Statistics and Data Analysis*, **61**, 148 – 157.

Sáez-Castillo, A.J. and Conde-Sánchez, A. (2016). Detecting over- and under-dispersion in zero inflated data with the hyper-Poisson regression model. *Statistical Papers.* In press.

TABLE 1. Coefficient estimates and standard errors (in parentheses) of ZIhP fitted model. In bold statistical significance at 5% level.

Variable	$\hat{\boldsymbol{\nu}}$	$\hat{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\delta}}$
Sex	-0.4711 (**0.1742**)	-0.0392 (0.0877)	
Age $\times 10^{-2}$	0.6591 (3.4422)	1.1256 (1.6921)	
Age$^2 \times 10^{-4}$	-1.7767 (3.7997)	-0.8670 (1.8218)	
Income	-0.0636 (0.2699)	-0.0868 (0.1434)	
Levyplus	-0.3309 (0.2092)	0.0646 (0.1203)	
Freepoor	0.4178 (0.5773)	-0.1852 (0.3211)	
Freerepat	-0.8705 (**0.2937**)	-0.0321 (0.1451)	
Illness	-0.5696 (**0.1108**)	-0.0327 (0.0361)	0.3128 (**0.1449**)
Activity days	-1.3631 (**0.2588**)	0.0760 (**0.0119**)	0.8672 (**0.2579**)
G. health q. score	-0.1070 (**0.0388**)	0.0220 (0.0143)	
Chronic cond. 1	-0.1832 (0.1970)	-0.0675 (0.1097)	
Chronic cond. 2	-0.3433 (0.2859)	0.0267 (0.1271)	
Constant	2.6041 (**0.6068**)	-0.6969 (**0.3215**)	-1.5493 (**0.5042**)

Sellers, K. F. and Shmueli, G. (2013). Data Dispersion: Now You See It... Now You Don't. *Communications in Statistics - Theory and Methods*, **42 (17)**, 3134–3147.

Staub, K. E. and Winkelmann, R. (2013). Consistent estimation of zero inflated count models. *Health Economics*, **22 (6)**, 673–686.

Modelling of the unemployment duration in the Czech Republic with the use of a finite mixture model

Ivana Malá

[1] University of Economics in Prague, Czech Republic

E-mail for correspondence: malai@vse.cz

Abstract: Unemployment belongs to the most serious economic and social problems of developed countries. The unemployment duration in the Czech Republic in 2008, 2010 and 2014 is analysed with the use of the methods of the survival analysis as an time-to-event random variable. A finite mixture of lognormal distributions is used to describe an overall distribution as well as the distribution of components given by education of the unemployed. Data from the Labour Force Sample Survey that is performed by the Czech Statistical Office are used for the analysis. In the questionnaire that is used for the survey the unemployment duration is given only in intervals. Data are treated as right censored and interval censored, exact values of the unemployment duration are not included in the analysed dataset. The strong positive effect of education on the duration of unemployment is quantified for all analysed years. An increase in unemployment duration is described for the period of the 2008-2012 economic crisis (2010) with respect to the periods before (2008) and after (2014) the crisis.

Keywords: censored data; finite mixture of distributions; unemployment.

1 Finite mixture model

Data from the Labour Force Sample Survey (LFSS) that is performed quarterly by the Czech Statistical Office (LFSS, 2015) are used in order to model the distribution of the unemployment duration. The households (statistical units) in the LFSS survey form a rotating panel, one fifth of the sample rotates quarterly and none of the households is followed for more than one year. During the interviews of the survey no exact durations of unemployment are recorded and respondents give the duration of unemployment in given intervals (in months) 0-1, 1-3, 3-6, 6-2, 12-18, 18-24, 24-48 and over 48 months. It means that all data are censored (incomplete), for an unemployed who found a job during the year in the survey

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

(four consequtive quarters, five visits) the observed duration is interval censored; for those who did not find a new job we know that the duration of unemployment is longer than the given interval and the datum is treated as right censored in the left limit of the recorded interval. For the analysis the durations shorter than two years were used with 4 levels for education of the unemployed (basic (ISCED-97 0, 1), secondary without baccalaureate (ISCED-97 2), secondary (ISCED-97 3, 4), tertiary (ISCED-97 5, 6)). There were found 2,893 (new job (interval censored data) 1,127) eligible unemployed people in 2008, 4,753 (1,501) in 2010 and 2,844 (1,418) in 2014. For T an unemployment duration in the Czech Republic, separate models were constructed for three analysed years. A mixture of four lognormal distributions is given (for a selected year, no subscript for a year is used, but the parameters are year-specific) by the density

$$f(t;\boldsymbol{\psi}) = \sum_{j=1}^{4} \pi_j f(t;\mu_j,\sigma_j),$$

where for $j = 1,...,4$ $f(t;\mu_j,\sigma_j)$ are component lognormal densities (from the basic education ($j = 1$) to tertiary ($j = 4$)) and π is a vector of weights of the components in the mixture. The vector of unknown parameters in the model $\boldsymbol{\psi}$ consists of the parameters (8 parameters) of component distributions $\mu_j,\sigma_j, j = 1,...,4$ and 3 free parameters π_j. The maximum likelihood estimates of parameters (Lawless (2003)) were evaluated in the program R, the package Survival was used, Therneau (2015).

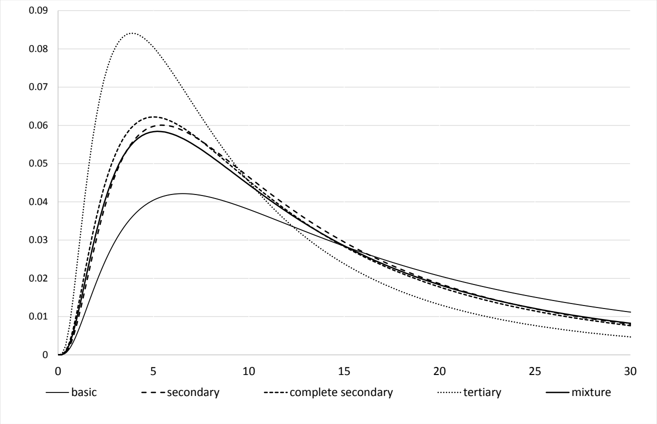


FIGURE 1. Fitted distributions for 2008.

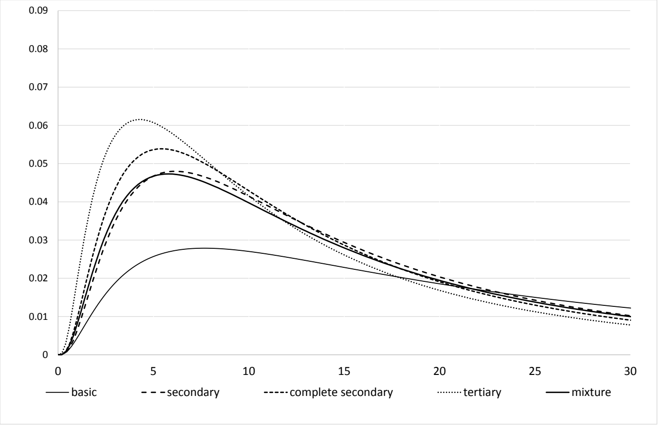


FIGURE 2. Fitted distributions for 2010.

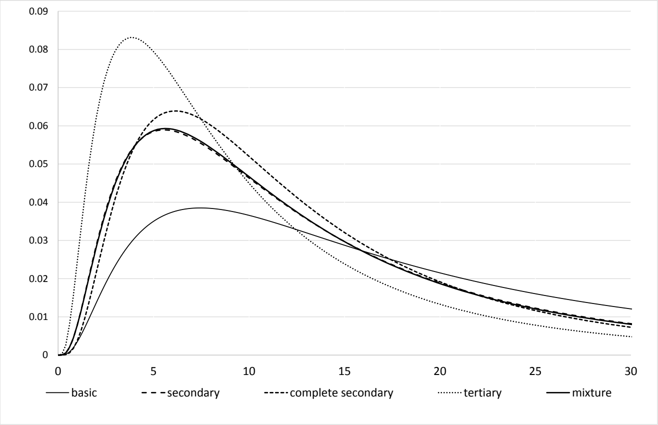


FIGURE 3. Fitted distributions for 2014.

2 Results

The mean age of the unemployed was 36.6-37.5 years (standard deviations 12.9-13.1), with no differences in age between men and women. The difference in age between the unemployed with aned without a new job was about 2 years in all the analysed periods. For all analysed years the highest values of estimated parameters μ, σ were found for basic education; parameters μ decrease with increasing education, variance parameters σ are comparable through the components (the use of one parameter for the variance decreases number of estimated parameters and improves the numerical performance, but based on the results of simulations the component specific variances are used). In 2010 both estimated parameters are greater then in 2008 or 2014, reflecting the length of unemployment during the economic crisis. Estimated component densities are given in the Figure 1-Figure 3 together with the density of the estimated mixture. The densities for tertiary and basic education are separated, densities for secondary and complete secondary education are similar and the density of the mixture coincides with them. One lognormal distribution was fitted into data and similar characteristics were obtained for the distribution of the duration of unemployment, however according to AIC criterion the mixture model provides better fit to data.

In order to compare components in years the characteristics of the level and the variability were evaluated from the estimated distributions (and quantile characteristics were prefered to the moment characteristics due to the heavy tails and positive skewness of the analysed distributions). The strong positive impact of education on the unemployment duration is clear from the figures as well as from the characteristics of the level. Difference in estimated median unemployment duration between basic and tertiary education is 7.7 (2008), 11.6 (2010) and 9.1 months (2014) with the population level 11.4, 14.3 and 11.3 months. For the quartile deviation these differences are 5.8, 11.0 and 6.4 months, for the whole population the estimated quartile deviations are 7.3, 9.5 and 6.6 months.

The hazard functions were estimated for all the models. All estimated hazard functions have one maximum, for the mixture at 9-10 months with maximums for the tertiary education at 7-8 months and 11-13 months for basic education.

References

Lawless, J. F. (2003). *Statistical models and methods for lifetime data.* 2nd ed. Hoboken: John Wiley & Sons.

LFSS. (2015). Labour Force Sample Survey. Czech Statistical Office. url = http://www.czso.cz/csu/czso/zam_vsps/.

Therneau, T. (2015). Package for Survival Analysis in S. version 2.38. url = http://CRAN.R-project.org/package=survival.

Presmoothed Landmark estimators of the transition probabilities

Luís Meira-Machado[1]

[1] Centre of Mathematics & Department of Mathematics and Applications, University of Minho, Campus de Azurem, 4800-058 Guimarães, Portugal.

E-mail for correspondence: lmachado@math.uminho.pt

Abstract: Multi-state models can be successfully used to model complicated event history data, for example, describing stages in the disease progression of a patient. In these models one important goal is the estimation of the transition probabilities since they allow for long term prediction of the process. There have been several recent contributions for the estimation of the transition probabilities. Recently, de Uña-Álvarez and Meira-Machado (2015) proposed new estimators for these quantities, and their superiority with respect to the competing estimators has been proved in situations in which the Markov condition is violated. In this paper, we propose a modification of the estimator proposed by de Uña-Álvarez and Meira-Machado based on presmoothing. Simulations show that the presmoothed estimators may be much more efficient than the completely nonparametric estimator.

Keywords: Kaplan-Meier; Multi-state model; Nonparametric estimation; Transition probabilities.

1 Introduction

In many medical studies individuals can experience several events across a follow-up study. Analysis of such studies can be successfully performed using a multi-state model (Meira-Machado et al., 2009). This paper introduces and studies a feasible estimation method for the transition probabilities in a progressive multi-state model.

Fully non-Markov estimators for the transition probabilities were introduced for the first time in Meira-Machado et al. (2006). Recently, this problem has been reviewed, and new sets of estimators have been proposed (de Uña-Álvarez and Meira-Machado, 2015). This method proceeds by considering specific subsets of individuals (namely, those observed to be in a given state at a pre-specified time

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

point) for which the ordinary Kaplan-Meier survival function leads to a consistent estimator of the target. Superiority with respect to the competing estimators has been proved.

A multi-state model is a stochastic process $(X(t), t \in \mathcal{T})$ with a finite state space, where $X(t)$ represents the state occupied by the process at time t. For simplicity, in this paper we assume the progressive illness-death model and we assume that all the subjects are in State 1 at time $t = 0$. The illness-death model describes the dynamics of healthy subjects (State 1) who may move to an intermediate "diseased" state (State 2) before entering into a terminal absorbing state (State 3). Many longitudinal medical data with multiple endpoints can be reduced to this structure.

The illness-death model is characterized by the joint distribution of (Z, T), where Z is the sojourn time in the initial state 1 and T is the total survival time. Both Z and T are observed subject to a random univariate censoring C assumed to be independent of (Z, T). Due to censoring, rather than (Z, T) we observe $(\widetilde{Z}, \widetilde{T}, \Delta_1, \Delta_2)$ where $\widetilde{Z} = \min(Z, C)$, $\Delta_1 = I(Z \leq C)$, $\widetilde{T} = \min(T, C)$, $\Delta_2 = I(T \leq C)$, where $I(\cdot)$ is the indicator function. The target is each of the five transition probabilities $p_{ij}(s, t) = P(X(t) = j \mid X(s) = i)$, where $1 \leq i \leq j \leq 3$ and $s < t$ are two pre-specified time points.

2 Estimators

The transition probabilities are functions involving expectations of particular transformations of the pair (Z, T). In practice, we only need to estimate three transition probabilities since the others can be expressed from these ones; namely,

$$\begin{aligned} p_{11}(s,t) &= \frac{E\left[I(Z > t)\right]}{E\left[I(Z > s)\right]}, \qquad p_{13}(s,t) = \frac{E\left[I(s < Z, T \leq t)\right]}{E\left[I(Z > s)\right]}, \\ p_{23}(s,t) &= \frac{E\left[I(Z \leq s < T \leq t)\right]}{E\left[I(Z \leq s < T)\right]}. \end{aligned}$$

Because of space limitation we will focus on the transition probability $p_{13}(s, t) = P(T \leq t \mid Z > s)$. Given the time point s, to estimate this quantity, the analysis can be restricted to the individuals with an observed first event time greater than s. This is known as the landmark approach (van Houwelingen et al. 2007). The corresponding estimator (KMW) is given by $\widehat{p}_{13}(s, t) = 1 - \widehat{S}_T^{(s)}(t)$ where $\widehat{S}_T^{(s)}(t)$ denote the Kaplan-Meier estimator computed from the given sub sample. Similarly, the transition probability $p_{23}(s, t) = P(T \leq t \mid Z \leq s < T)$ can be estimated by considering specific subsets of individuals (namely, those observed to be in a state 2 at a pre-specified time point s, i.e. those for which $Z \leq s < T$) for which the ordinary Kaplan-Meier survival function leads to a consistent estimator (see de Uña-Álvarez and Meira-Machado (2015) for further details).

The standard error of the estimators introduced by de Uña-Álvarez and Meira-Machado may be large when censoring is heavy, particularly with a small sample size. Interestingly, the variance of this estimator may be reduced by presmoothing. This 'presmoothing' is obtained by replacing the censoring indicator variables in the expression of the Kaplan-Meier weights (used by the authors), by a smooth fit. This preliminary smoothing may be based on a certain parametric family such as the logistic, or on a nonparametric estimator of the binary regression curve. The (semiparametric) Kaplan-Meier Presmoothed Weighted estimator (KMPW) is given by $\widehat{p}^{\star}_{13}(s,t) = 1-\widehat{S}_T^{(s\star)}(t)$ where $\widehat{S}_T^{(s\star)}(t)$ denotes the presmoothed Kaplan-Meier estimator in the same sub sample.

Note that, unlike the estimator by de Uña-Álvarez and Meira-Machado (2015), the semiparametric (presmoothed) estimator can attach positive mass to pair of event times with censored total time. However, both estimators attach a zero weight to pairs of event times for which the first event time is censored. In the limit case of no presmoothing, the two estimators are equivalent.

3 Simulation study

In this section we investigate the performance of the proposed estimators through simulations. The simulated scenario is the same as that described in Amorim et al. (2011). To compare the performance of the methods we compute the mean square error (MSE), bias and standard deviation (SD). For completeness we also included the estimator by Meira-Machado et al. (2006).

Figure 1 depicts the boxplots of the estimated MSE over 1000 simulated datasets. From this plot it can be seen that with exception of the (LIDA) estimator by Meira-Machado et al. (2006) the remaining two estimators (KMW and KMPW) perform well, approaching their targets as the sample size increases. Besides, simulation results also reveal that the new proposal perform favorably when compared with the competing methods. Our simulation results reveal relative benefits of presmoothing in the heavily censored scenarios or small sample sizes.

Acknowledgments: This project was funded by FEDER Funds through “Programa Operacional Factores de Competitividade - COMPETE” and by Portuguese Funds through FCT - “Fundação para a Ciência e a Tecnologia”, in the form of grant PEst-OE/MAT/UI0013/2014.

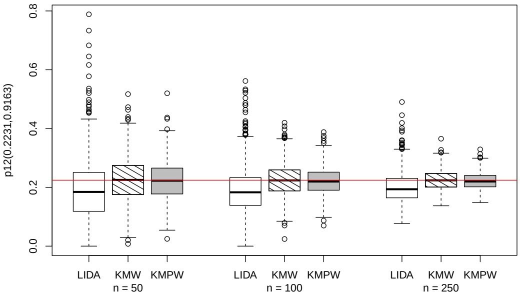
References

Amorim, A. P., de Uña-Álvarez, J., and Meira-Machado, L. (2011). Presmoothing the transition probabilities in the illness-death model. *Statistics & Probability Letters*, **81**, 797 – 806.

de Uña-Álvarez, J., and Meira-Machado, L. (2015). Nonparametric estimation of transition probabilities in the non-Markov illness-death model: a comparative study *Biometrics*, **71(2)**, 364 – 375.

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

Meira-Machado, L., de Uña-Álvarez, J., and Cadarso-Suárez, C. (2006). Nonparametric estimation of transition probabilities in a non-Markov illness-death model. *Lifetime Data Analysis*, **12**, 325 – 344.



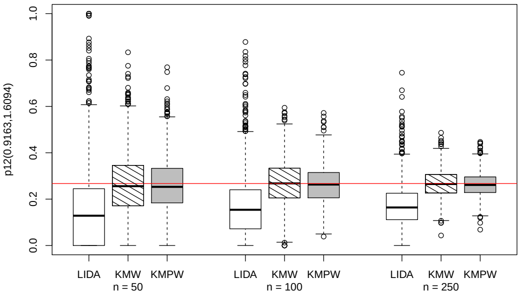


FIGURE 1. Mean square error for the three estimators.

Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C., and Andersen, P.K. (2009). Multi-state models for the analysis of time to event data *Statistical Methods in Medical Research*, **18**, 195 – 222.

van Houwelingen, H.C. (2007). Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, **34**, 70 – 85.

Modeling market-shares or compositional data: An application to the automobile market

Joanna Morais[1], Christine Thomas-Agnan[1], Michel Simioni[2]

[1] Toulouse School of Economics, France
[2] INRA, UMR 1110 MOISA, Montpellier, France

E-mail for correspondence: `joanna.morais@ut-capitole.fr`

Abstract: To analyze the impact of marketing investment on sales in the automobile market, the competitive context cannot be ignored. Thus market-shares and shares-of-voice are of interest. This contribution aims to present and compare statistical modeling methods adapted for shares or proportions, which are characterized by the following constraints: positivity and sum equal to 1. Two major approaches address this question: market-share models from the econometric marketing literature and compositional data analysis from the statistical literature. The common point between the two is to use log-ratio transformations in order to model shares accounting for their constraints. The differences are mostly coming from the assumptions made on the distribution of the data itself or on the error terms of the models.

Keywords: Market-share; Compositional data; Automobile.

1 Introduction

This paper aims to present statistical modeling methods adapted for shares or proportions, which are characterized by the following constraints: they are positive and sum up to 1. By definition shares are "compositional data": a composition is a vector of parts of some whole which carries relative information. For a composition of D parts, if $D-1$ parts are known the D^{th} is simply 1 minus the sum of the $D-1$ parts. Thus, a D-composition could be represented with a $(D-1)$-dimensional real Euclidean vector space structure. Because of these constraints, classical regression models cannot be used directly. The well-known *ceteris paribus* is unusable to interpret models based on compositional data because when one share changes, the others change too.

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

Two major branches of the statistical literature take into account the constraints of this type of data: market-share models coming from the quantitative marketing applications, and compositional models coming from CODA (Compositional Data Analysis).

2 Market-share models

Market-share models were developed in the 80's, mainly by Nakanishi & Cooper (1988). The aim is to model market-shares of D brands using their marketing factors (price, advertising) as explanatory variables.

These models are aggregated versions of discrete choice models like the MNL model (Multinomial Logit model). The concept of "attractivity" of a brand is central in this literature, and is comparable to the "utility" concept in discrete choice models. The specification of the attractivity of brand j is a multiplicative expression of the explanatory variables describing brand j. The market-share of brand j is defined as its relative attractivity compared to competitors, i.e. as its attractivity divided by the sum of attractivities of all the brands of the market. This explains why they are often called MCI models (Multiplicative Competitive Interaction models).

The estimation of MCI models is usually made on the log-linearized shares with OLS or GLS. But one can prove that under certain assumptions, a maximum likelihood estimation could be done. This literature is concerned by explaining shares so the focus is on the case of a dependent compositional variable whereas we will see that in CODA, the compositional nature of the variable can also be for the independent variables.

3 Compositional models

Compositional data analysis was developed first in the 80's by John Aitchison (1986). Since the 90's, a group of researchers from the University of Girona (V. Pawlowsky-Glahn, J.J. Egozcue, C. Barcelo, J.A. Martin Fernandez) is particularly active in the domain and has developed a large mathematical framework for this literature.

First applications were made on geological data, with the objective to analyze the composition of a rock sample in terms of the relative presence of different chemical elements. More generally, CODA aims to analyze relative information between the components (parts) of a composition where the total of the components is not relevant or is not of interest.

A composition of D parts lies in the simplex $\mathcal{S}^D$. The suited geometry for compositions is the Aitchison geometry (or simplicial geometry), not the Euclidean geometry. Aitchison geometry defines the perturbation operation $\oplus$, the powering operation $\odot$ and the inner product $\langle ., . \rangle_a$ inside the simplex. Compositions can be transformed in coordinates by a log-ratio transformation in order to be represented in a $\mathbb{R}^{D-1}$ Euclidean space. Then, classical methods suited for data in the Euclidean space can be used on coordinates. Three main transformations are developed: the ALR (additive log-ratio), the CLR (centered log-ratio) and the ILR (isometric log-ratio) transformations, each of them having specific advantages.

CODA regroups different tools for analyzing compositions: graphical tools (ternary diagrams, biplot of coordinates), analytic tools (compositional PCA) and compositional regression models. Compositional regression models are of different types depending on whether the response variable and/or the explanatory variables are compositional. The estimation is made on coordinates (after log-ratio transformation), usually with the OLS method.

4 Application

These methods are applied to an automobile market data set containing for each vehicle, by month, the sales volume, the catalog price, the media investments by canal (TV, Press, Radio, Outdoor, Digital, Cinema), and attributes of the vehicle (brand, segment, age on the market). The main objective is to understand the impact of media investments on market-shares in terms of sales volume (the response variable is compositional) controlling for other factors (attributes and price).

Moreover, two other compositional models are tested. The 1st model uses a composition as explanatory variable (the composition of media expenses by canal for each vehicle) while the response variable (the sales for each vehicle) and the others explanatory variables (price and attributes for each vehicle) are in volume. The 2nd model considers compositions in the response variable side (the market-shares of all vehicles) and in the explanatory variable side (the composition of relative media expenses of all vehicles, so-called shares-of-voice in marketing).

In each model specification the interest is on the marginal impact of each canal of media and on the interactions between the different canals. The final objective is to do some recommendations for the car manufacturer on the total amount of media to invest and on its distribution among the canals.

5 Comparison and results

A comparison of these two model families is done in terms of properties of results, assumptions on the distribution of error terms, and interpretation of fitted coefficients. It turns out that the two methods have benefits and drawbacks, and can benefit to each other.

On the one hand, we prove that the IIA (Independence from Irrelevant Alternatives) property of discrete choice models (and then of market-share models) is equivalent to the sub-compositional coherence of CODA models. Whereas the IIA property has been often criticized (because it turns out to be often unrealistic in practice) leading to the development of more flexible models in discrete choice models (nested logit, GEV) with their counterpart in market-share models, the sub-compositional coherence is apparently never questionned in the CODA literature.

On the other hand, CODA proposes the ILR transformation which is a projection of compositional data using an orthonormal basis in the simplex. ILR has good mathematical properties and permits to use a large number of statistical tools, contrary to ALR and CLR. However, one can show that the transformations used in market-share models correspond to ALR or CLR, but ILR is never used.

Moreover, assumptions on covariance structure between the components with MNL specification are less flexible (negative covariance is imposed) than in the CODA framework (positive covariance is allowed between certain components). Finally, concerning the results of the modeling, the two methods will lead to different interpretations and complementary information depending on the log-ratio transformation used.

Acknowledgments: Special Thanks to Renault, especially Marc-Henri Ambroise, Philippe Devailly and Matthias Guegan who give me access to valuable data and shares with me their business expertise. Thanks a lot to Herve Tranger from the market research institute BVA, my thesis tutor, and to Christine Thomas-Agnan and Michel Simioni, my thesis advisors, for their day-to-day commitment.

References

Nakanishi, M. and Cooper, L.G. (1982). Simplified Estimation Procedures for MCI Models. *Marketing Science*, **1**, 314 – 322.

Cooper, L.G. and Nakanishi, M. (1988). *Market-Share Analysis: Evaluating Competitive Marketing Effectiveness*. Springer.

Haaf, C. G., Michalek, J., Morrow, W. R. and Liu, Y. (2014). Sensitivity of Vehicle Market Share Predictions to Discrete Choice Model Specification. *Journal of Mechanical Design*, **136**.

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman and Hall.

Van Den Boogaart, K.G. and Tolosana-Delgado, R. (2013). *Analysing Compositional Data with R*. Springer.

Pawlowsky-Glahn, V., Egozcue, J.J. and Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data*. John Wiley & Sons.

Probability of causation: Bounds and identification for partial and complete mediation analysis

Rossella Murtas[1], Alexander Philip Dawid[2], Monica Musio[1]

[1] PhD student at the University of Cagliari, Italy
[2] Leverhulme Emeritus Fellow, University of Cambridge, UK
[3] Professor in Statistics, University of Cagliari, Italy

E-mail for correspondence: ro.murtas@gmail.com

Abstract: An individual has been subjected to some exposure and has developed some outcome. Using data on similar individuals, we wish to evaluate, for this case, the probability that the outcome was in fact caused by the exposure. Even with the best possible experimental data on exposure and outcome, we typically can not identify this " probability of causation" exactly, but we can provide information in the form of bounds for it. Here, using the potential outcome framework, we propose new bounds for the case that a third variable mediates partially or completely the effect of the exposure on the outcome.

Keywords: Probability Of Causation; Mediation Analysis; Potential Outcomes.

1 Introduction

A typical causal question can be categorized into two main classes: about the causes of observed effects, or about the effects of applied causes. Let us consider the following example: an individual, called Ann, might be subjected to some exposure X, and might develop some outcome Y. We will denote by $X_A \in \{0,1\}$ the value of Ann's exposure (coded as 1 if she takes the drug) and by $Y_A \in \{0,1\}$ the value of Ann's outcome (coded as 1 if she dies). Questions on the effects of causes, named "EoC", are widely known in literature as for example by Randomized clinical trials. In the EoC framework we would be interested in asking: "What would happen to Ann if she were (were not) to take the drug?". On the other hand, questions on the causes of observed effects, "CoE", are common in a Court of Law, when we want to assess legal responsibility. For example, let us suppose that Ann has developed the outcome after being exposed, a typical question will be "Knowing that Ann did take the drug and passed away, how likely

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

is it that she would not have died if she had not taken the drug?". In this paper we will discuss causality from the CoE perspective, invoking the potential outcome framework. Definition of CoE causal effects invokes the *Probability of Causation* Pearl (1999) and Dawid (2011) $\mathrm{PC}_A = \mathrm{P}_A(Y_A(0) = 0 \mid X_A = 1, Y_A(1) = 1)$ where P_A denotes the probability distribution over attributes of Ann and $Y(x)$ is the hypothetical value of Y that would arise if X was set to x. Note that this expression involves the bivariate distribution of the pair $\mathbf{Y} = (Y(0), Y(1))$ of potential outcomes. Whenever the probability of causation exceeds 50%, in a civil court, this is considered as preponderance of evidence because causation is " more probable than not".

2 Starting Point: Simple Analysis

In this section we discuss the simple situation in which we have information, from a hypothetical randomized experimental study (such that $X_i \perp\!\!\!\perp \mathbf{Y}_i$ for a subject i in the experimental population) that tested the same drug taken by Ann such that $P_1 = \mathrm{P}(Y = 1 \mid X \leftarrow 1) = 0.30$ and $P_0 = \mathrm{P}(Y = 1 \mid X \leftarrow 0) = 0.12$. This information alone is not sufficient to infer causality in Ann's case. We need to further assume that the fact of Ann's exposure, X_A, is independent of her potential responses $\mathbf{Y}_A$, that is $X_A \perp\!\!\!\perp \mathbf{Y}_A$, and that Ann is exchangeable with the individuals in the experiment. On account of this and exchangeability, the PC_A reduces to $\mathrm{PC}_A = \mathrm{P}(Y(0) = 0 \mid Y(1) = 1)$. However, we can never observe the joint event $(Y(0) = 0; Y(1) = 1)$, since at least one event must be counterfactual. But even without making any assumptions about this dependence, we can derive the following inequalities, Dawid *et al.* (2015):

$$1 - \frac{1}{\mathrm{RR}} \leq \mathrm{PC}_A \leq \frac{\mathrm{P}(Y = 0 \mid X \leftarrow 0)}{\mathrm{P}(Y = 1 \mid X \leftarrow 1)} \tag{1}$$

where $\mathrm{RR} = \mathrm{P}(Y = 1 \mid X \leftarrow 1)/\mathrm{P}(Y = 1 \mid X \leftarrow 0)$ is the *experimental risk ratio* between exposed and unexposed. Since, in the experimental population, the exposed are 2.5 times as likely to die as the unexposed $(\mathrm{RR} = 30/12 = 2.5)$, we have enough confidence to infer causality in Ann's case, given that $0.60 \leq \mathrm{PC}_A \leq 1$.

3 Bounds in Mediation Analysis

In this Section we present a novel analysis to bound the Probability of Causation for a case where a third variable, M, is involved in the causal pathway between the exposure X and the outcome Y and plays the role of mediator. We shall be interested in the case that M is observed in the experimental data but is not observed for Ann, and see how this additional experimental evidence can be used to refine the bounds on PC_A.

First we consider the case of complete mediation, Dawid *et al.* (2016). Using counterfactual notation, we denote by $M(x)$ the potential value of M for $X \leftarrow x$, and by $Y^*(m)$ the potential value of Y for $M \leftarrow m$. Then

$Y(x) := Y^*\{M(x)\}$. Assuming no confounding for the exposure-mediator and mediator-outcome relationship, the causal pathway will be blocked after adjustment for M (Markov property $Y \perp\!\!\!\perp X|M$). The assumed mutual independence implies the following upper bounds for the probability of causation in the case of complete mediation: $\mathrm{PC}_A \leq \mathit{Num}/\mathrm{P}(Y = 1 \mid X \leftarrow 1)$, while the lower bound remains unchanged from that of the simple analysis of X on Y in Eq. (1). For the upper bound's numerator, *Num*, one has to consider various scenarios according to different choices of the estimable marginal probabilities in Table 1.

TABLE 1. Upper Bound's Numerator for PC$_A$ in Complete Mediation Anlaysis

	$a \leq b$	$a > b$
$c \leq d$	$a \cdot c + (1-d)(1-b)$	$b \cdot c + (1-d)(1-a)$
$c > d$	$a \cdot d + (1-c)(1-b)$	$b \cdot d + (1-a)(1-c)$

In Table 1, $a = P(M(0) = 0)$, $b = P(M(1) = 1)$, $c = P(Y^*(0) = 0)$ and $d = P(Y^*(1) = 1)$. Given the no-confounding assumptions, these are all estimable probabilities.

For the case of partial mediation, we introduce: $Y^*(x, m)$, the potential value of the outcome after setting both exposure and mediator, so that now $Y(x) = Y^*(x, M(x))$. Let us consider the following assumptions (named **(A)**): $Y^*(x, m) \perp\!\!\!\perp (M(0), M(1))|X$; $Y^*(x, m) \perp\!\!\!\perp X$ that is no $X - Y$ confounding and $M(x) \perp\!\!\!\perp X$ that is no $X - M$ confounding. Note that assumption $Y^*(x, m) \perp\!\!\!\perp (M(0), M(1))|X$ implies both $Y^*(x, m) \perp\!\!\!\perp M(0)|X$ and $Y^*(x, m) \perp\!\!\!\perp M(1)|X$, that is no $M - Y$ confounding. If Ann is exchangeable with the individuals in the experiment
$\mathrm{PC}_A = \mathrm{P}(Y(0) = 0,\ Y(1) = 1 \mid X = 1)/\mathrm{P}(Y(1) = 1 \mid X = 1)$.
The numerator involves a bivariate distribution of counterfactual outcomes. Using assumptions **(A)** and and the inequality $P(A \cap B) \leq \min\{P(A), P(B)\}$, we can obtain an upper bound for PC$_A$ considering these 64 combinations

$$\begin{aligned}
&\mathrm{P}(Y(0) = 0,\ Y(1) = 1|X = 1) \leq \\
&min\{\mathrm{P}(Y^*(0,0) = 0), \mathrm{P}(Y^*(1,0) = 1)\} \cdot min\{\mathrm{P}(M(0) = 0), \mathrm{P}(M(1) = 0)\} && (2)\\
&+ min\{\mathrm{P}(Y^*(0,0) = 0), \mathrm{P}(Y^*(1,1) = 1)\} \cdot min\{\mathrm{P}(M(0) = 0), \mathrm{P}(M(1) = 1)\} && (3)\\
&+ min\{\mathrm{P}(Y^*(0,1) = 0), \mathrm{P}(Y^*(1,0) = 1)\} \cdot min\{\mathrm{P}(M(0) = 1), \mathrm{P}(M(1) = 0)\} && (4)\\
&+ min\{\mathrm{P}(Y^*(0,1) = 0), \mathrm{P}(Y^*(1,1) = 1)\} \cdot min\{\mathrm{P}(M(0) = 1), \mathrm{P}(M(1) = 1)\} && (5)
\end{aligned}$$

It can be proved that the lower bound does not change. Assumptions **A** will be enough to estimate the lower and the upper bounds from the data.

4 Comparisons and conclusions

The numerator of the upper bound of PC_A in the simple analysis framework (1), which ignores the mediator, may be written as

$$\min\{\mathrm{P}(Y^*(0,0)=0)\mathrm{P}(M(0)=0)+\mathrm{P}(Y^*(0,1)=0)\mathrm{P}(M(0)=1), \mathrm{P}(Y^*(1,0)=1)\mathrm{P}(M(1)=0) + \mathrm{P}(Y^*(1,1)=1)\mathrm{P}(M(1)=1)\} = \min\{\alpha+\beta, \gamma+\delta\}. \quad (6)$$

We can see that both (2) and (3) are smaller than or equal to α, while both (4) and (5) are smaller than or equal to β. Thus, the upper bound not accounting for the mediator, could be larger or smaller than that obtained considering the partial mediation mechanism. On the other hand, it can be proved that the bounds found for the case of complete mediation are never larger than for the simple analysis of X on Y.

In conclusion, the important implications of PC_A in real cases encourage the researcher to focus on studying methods capable of producing more precise bounds. Here we have proposed a novel analysis to bound the PC_A when a mediator lies on a pathway between exposure and outcome.

References

Pearl, Judea (1999). Probabilities of causation: three counterfactual interpretations and their identification. *Synthese*, **121 (1-2)**, 93 – 149.

Dawid, A. Philip (2011). The role of scientific and statistical evidence in assessing causality. In *Perspectives on Causation*, Oxford: Hart Publishing, 133 – 147.

Dawid, A. P, and Murtas, R. and Musio, M. (2016). Bounding the Probability of Causation in Mediation Analysis. In the Springer Book *Selected Papers of the 47th Scientific Meeting of the Italian Statistical Society*, in press. Editors: T. Di Battista, E. Moreno, W. Racugno. arXiv preprint arXiv:1411.2636.

Dawid, A. P, and Musio, M. and Fienberg, S. (2015). From statistical evidence to evidence of causality. *Bayesian Analysis*, Advance Publication, 26 August 2015. DOI:10.1214/15-BA968

Comparison of beta-binomial regression approaches to analyze health related quality of life data

Josu Najera-Zuloaga[1], Dae-Jin Lee[1], Inmaculada Arostegui[12]

[1] Basque Center for Applied Mathematics, Bilbao, Spain
[2] University of the Basque Country, Bilbao, Spain

E-mail for correspondence: `jnajera@bcamath.org`

Abstract: Health related quality of life (HRQoL) has become an increasingly important indicator of health status in clinical trials and epidemiological research. It has been stated that beta-binomial distribution is a good option to fit this type of outcomes. The goal of HRQoL analysis use to be the measurement of the effect of patients' and disease's characteristics have on HRQoL of patients. This work is motivated by the application in a real data with HRQoL observations of chronic obstructive pulmonary disease (COPD) patients in which two regression models based on the beta-binomial distribution yields contradictory results: i) Beta-binomial distribution with a logistic link and ii) hierarchical likelihood approach (HGLM). None of the existing literature in the analysis of HRQoL survey data has performed a comparison of both approaches in terms of adequacy. In this work we present a detailed comparison by a simulation study and propose the best approach in terms of parameter significance and effect to deal with HRQoL outcomes.

Keywords: Beta-binomial regression; health related quality of life; chronic obstructive pulmonary disease; Hierarchical GLM.

1 Introduction

Health-related quality of life (HRQoL) has become an important measure of health status as it provides patients information in a standardized, comparable and objective way. The relationships between HRQoL and risk factors can help to evaluate medical care results, especially in chronic diseases. One of the most widely used generic instruments for measuring HRQoL is the Short Form-36 (SF-36) Health Survey.

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

It has been proved in the literature that the beta-binomial distribution is an adequate candidate to fit SF-36 survey data and a beta-binomial regression model has been proposed to perform HRQoL analysis. There are two different approaches in the literature in order to implement a regression model based on the beta-binomial distribution: (i) beta-binomial distribution with a logistic link (Forcina and Franconi, 1988) and (ii) Hierarchical Generalized Linear Model (HGLM) by Lee and Nelder (1996). However, none comparison between these two regression approaches has been addressed from a practical point of view.

The application of the two regression approaches in a real data containing HRQoL measurements of chronic obstructive pulmonary disease (COPD) patients yields to contradictory results in terms of covariate effect. Consequently, we have performed a comparison study concluding the best regression approach in terms of covariate effect significance and interpretation when dealing with HRQoL data.

2 Methodologies

The beta-binomial distribution consists of a finite sum of Bernoulli dependent variables whose probability parameter is random and follows a beta distribution with parameters α_1 and α_2. In general, if θ is the probability variable, y follows a beta-binomial distribution if

$$y|\theta \sim \text{Bin}(m, \theta) \quad \text{and} \quad \theta \sim \text{Beta}(\alpha_1, \alpha_2).$$

The marginal likelihood of this distribution can be explicitly calculated.

2.1 Logistic regression based on a beta-binomial distribution

Let $y_1, \ldots, y_n$ be a set of independent beta-binomial variables. Arostegui et al. (2007) proposed a reparameterization as $\alpha_{1i} = p_i/\phi$ and $\alpha_{2i} = (1 - p_i)/\phi$. Consequently, we have that

$$\text{E}[y_i] = np_i \qquad \text{and} \qquad \text{Var}[y_i] = np_i(1 - p_i)\left[1 + (m-1)\frac{\phi}{1+\phi}\right],$$

which allows the interpretation of p_i as the probability parameter. Forcina and Franconi (1988) proposed to link p_i through a logit function connecting it with some covariates. They proposed an iterative estimation method based on the maximum likelihood approach. We call this model *BBlogit*.

2.2 HGLM

Lee and Nelder (1996) introduced the concept of hierarchical generalized linear models (HGLMs), as a generalization of generalized linear mixed models (GLMMs) to the inclusion of non-gaussian random effects. Consequently, we can formulate the following model:

$$y_i|u \sim \text{Bin}(m, p_i) \quad \text{and} \quad u_j \sim \text{Beta}(1/(2\phi), 1/(2\phi))$$

where the linear predictor is

$$\text{logit}(p_i) = x_i'\beta + z_i'v$$

and $v = \text{logit}(u)$ are the random effects. The estimation in this model is done via h-likelihood and some adjusted profile likelihoods.

3 Application to real data: COPD study

COPD is a very common chronic disease, which is expected to increase in prevalence over the next years. Both regression approaches were applied to COPD data. Depending on the HRQoL dimensions results were completely different. In *role emotional* dimension the results were:

TABLE 1. Effect of explanatory variables in *role emotional* dimension measured by both beta-binomial regression approaches.

Role emotional	*hglm*			*BBlogit*		
	$\hat{\beta}$	SD($\hat{\beta}$)	p-value	$\hat{\beta}$	SD($\hat{\beta}$)	p-value
Anxiety						
Yes	-6.145	2.062	0.003	-1.649	0.226	<0.001
Dyspnea						
Mild	-2.600	5.229	0.619*	-0.614	0.418	0.142*
Modere	-3.981	5.080	0.434*	-1.379	0.413	<0.001
Severe	-5.603	5.496	0.309*	-2.048	0.467	<0.001
log(ϕ)	2.735	0.095	<0.001	0.668	0.150	<0.001

Different conclusions were obtained depending on the applied approach, which evidences de need of a simulation study to compare them.

4 Simulation study

We have performed 500 simulations of 100 observations of a beta-binomial dependent variable, which probability parameter was calculated by a simulated covariate and by fixing the regression parameters β_0 and β_1. The value of ϕ divides the simulation in different scenarios: bell-shaped ($\phi < 0.5$), flat-shaped ($\phi = 0.5$) and U-shaped ($\phi > 0.5$). Table 2 shows the results in U-shaped scenario ($\phi = 2$), the mean and standard deviation of the regression coefficients, the expected mean squared (EMS) and the percentage the covariate effect is statistically significant (PCSS) are shown.

Results show that *BBLogit* is the best performance not only in terms of regression parameters estimation and EMS, but also measuring the statistical significance of the covariate. We should emphasize that the *hglm* model only considers the covariate statistically significant in 22.6%.

TABLE 2. Comparison results from a simulation of 500 replicates with $\phi = 2$ and 100 observed responses, m parameter was fixed at $m = 20$.

True value	$\beta_0 = 1$		$\beta_1 = -0.3$		
Method	Mean(SD)	EMS	Mean(SD)	EMS	PCSS
hglm	3.192 (2.613)	11.630	-0.949	0.874 (0.673)	22.6%
BBlogit	0.789 (0.311)	0.141	-0.248	0.010 (0.086)	82.0%

5 Discussion

We have illustrated that the use of two approaches for beta-binomial regression analysis may lead to different interpretation of the effect of the covariates. HGLM approach considers the expectation of the random components u equal to 1/2, which, as the dispersion parameter increases, will not be maintained by the logit transformation, reaching non zero mean random effects and creating bias in the estimation. Little variations create great differences in the mean of the logit transformation, which increases the variance, and consequently, significance tests fail.

Hence, if the HRQoL analysis interest lies in the interpretation of the regression coefficients, we suggest the use of the *BBlogit* methodology.

Acknowledgments: This research was supported by the Basque Government through the BERC 2014-2017 and the IT-620-13 programs, by the Spanish Ministry of Economy and Competitiveness MINECO: BCAM Severo Ochoa excellence accreditation SEV-2013-0323.

References

Arostegui, A., Nuñez-Antón, V., and Quintana, J. M. (2007). Analysis of the short form-36 (SF-36): The beta-binomial distribution approach. *Statistics in Medicine*, **26**, 1318 – 1342.

Forcina, A. and Franconi, L. (1988). Regression analysis with the beta-binomial distribution. *Rivista si Statistica Applicata*, **21**.

Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B*, **58**, 619 – 678.

A new flexible distribution with unimodal and bimodal shapes

Luiz R. Nakamura[1], Robert A. Rigby[2], Dimitrios M. Stasinopoulos[2], Roseli A. Leandro[1], Cristian Villegas[1], Rodrigo R. Pescim[3]

[1] Departamento de Ciências Exatas, Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Brazil
[2] STORM Research Centre, London Metropolitan University, United Kingdom
[3] Departamento de Estatística, Universidade Estadual de Londrina, Brazil

E-mail for correspondence: `lrnakamura@usp.br`

Abstract: In this work we consider a new family of distributions called the Birnbaum-Saunders power distribution that presents a bimodal shape for certain values of parameters. We study its probability density function and present an application regarding to the waiting time between eruptions in a geyser in the USA comparing some of the new distributions.

Keywords: Bimodality; Birnbaum-Saunders distribution; Positive data.

1 Introduction

Let Z follow any distribution on the real line, denoted by $Z \sim \mathcal{D}(\theta)$, with parameter vector θ, and let

$$Y = \psi \left[\frac{Z}{2} + \sqrt{\left(\frac{Z}{2}\right)^2 + 1} \right]^{\xi}, \tag{1}$$

where $Y > 0$, then the distribution of Y is named here as the Birnbaum-Saunders power (BSP) distribution, where $\psi > 0$ is a scale parameter and $\xi > 0$ is a skewness parameter. For simplicity we will assume, from now on, that Z follows a distribution with up to four parameters, i.e., $\theta = (\mu, \sigma, \nu, \tau)^\top$, where $-\infty < \mu < \infty$ is the location parameter, $\sigma > 0$ is the scale parameter and ν and τ are usually parameters related to the tails

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

of the distribution of Z. However, after the transformation is performed, μ and σ are called non-centrality and shape parameters respectively. Hence, the resulting BSP distribution for Y has up to six parameters and will be denoted as $Y \sim BSP(\psi, \xi, \mu, \sigma, \nu, \tau)$. Note that, if $Z \sim N(0, \sigma^2)$ and $\xi = 2$ we have the standard Birnbaum-Saunders distribution (Birnbaum and Saunders, 1969). Moreover, if $\xi = 2$ and Z follows any symmetric distribution with location parameter $\mu = 0$ we have the generalised Birnbaum-Saunders distribution (Díaz-García and Leiva, 2005). Finally, the BSP probability density function can be written as

$$f_Y(y|\psi, \xi; \theta) = f_Z(z|\theta) \left| \frac{dz}{dy} \right|, \, y > 0,$$

where θ corresponds to the parameters inherited from the baseline distribution and

$$\frac{dz}{dy} = \frac{1}{y\xi} \left[\left(\frac{y}{\psi} \right)^{\frac{1}{\xi}} + \left(\frac{y}{\psi} \right)^{-\frac{1}{\xi}} \right].$$

The BSP distribution is potentially a very flexible distribution for $Y > 0$, depending on the flexibility of the distribution of Z, and can be unimodal or bimodal. When $\xi \rightarrow 0$ and σ is large, BSP distributions present bimodality.

2 Special cases of the BSP distribution

As explained in Section 1, Z can follow any distribution on the real line. In this work, we used four different baseline distributions for Z creating four different distributions: i) Birnbaum-Saunders power normal (BSPNO); ii) Birnbaum-Saunders power generalised t (BSPGT); and iii) Birnbaum-Saunders power Johnson S_u (BSPJSU); and iv) Birnbaum-Saunders power sinh-arcsinh (BSPSHASHo) distributions. See Stasinopoulos and Rigby (2007) for details of these baseline distributions.

3 Application

The data refers to the waiting time between eruptions for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA and it is available on `R` software. Table 1 displays the MLEs of the BSPNO, BSPGT, BSPJSU and BSPSHASHo parameters and their corresponding AIC value.

We can see that the BSPSHASHo distribution outperformed all the other distributions, since it produced the smallest value of AIC (2075.49). Plots of the fitted distributions are displayed in Figure 1.

Figure 2 displays the (normalised quantile) residuals from the fitted BSPSHASHo distribution. The true residuals have a standard normal distribution. Panel (a) is a plot of residuals against the case number (index), panels

TABLE 1. MLE of model parameters for the waiting time between eruptions.

Model	$\hat{\psi}$	$\hat{\xi}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\nu}$	$\hat{\tau}$	AIC
BSPSHASHo	66.69	0.09	5.62	3.74	-0.19	0.44	2075.49
BSPJSU	65.30	0.09	3.25	25.79	-0.19	0.88	2078.79
BSPGT	64.14	0.09	6.00	10.41	114.21	0.96	2081.08
BSPNO	62.55	0.13	3.43	6.72	–	–	2087.99

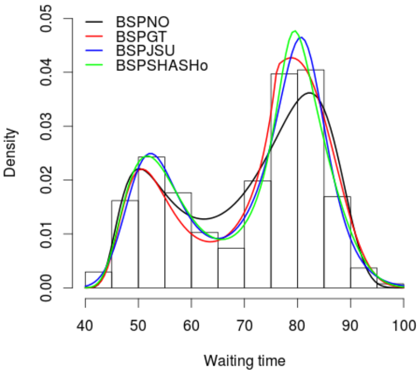


FIGURE 1. Comparison of the fitted distributions to the waiting time between eruptions.

(b) and (c) display a kernel density estimate and a normal QQ plot for the residuals, respectively, and panel (d) shows the worm plot (van Buuren and Fredriks, 2001) of the residuals. The residuals adequately follow a normal distribution and appear random. Moreover there are no problems in the worm plot.

4 Conclusion

We presented a new flexible bimodal family of distributions, called the BSP distribution and show its flexibility through a real data set application related to the waiting time between eruptions for the Old Faithful geyser in the USA.

Acknowledgments: The first author gratefully acknowledge grant from CAPES (Brazil) under the process number 99999.009857/2014-01.

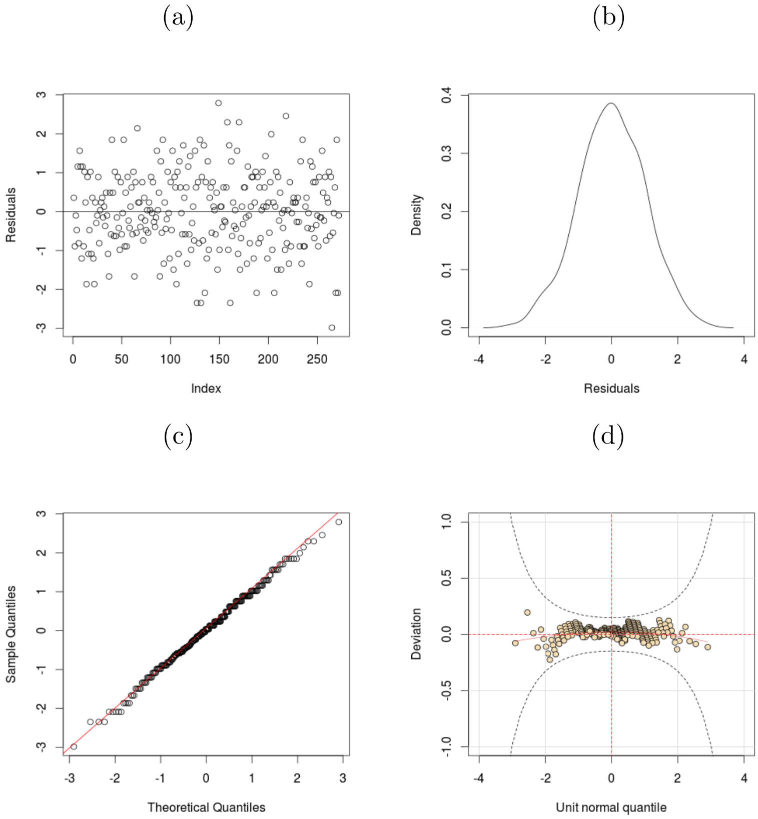


FIGURE 2. Residual plots for the fitted BSPSHASHo distribution.

References

Birnbaum, Z.W. and Saunders, S.C. (1969). A new family of life distributions. *Journal of Applied Probability*, **6**, 319 – 327.

van Buuren, S. and Fredriks, M. (2001). Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, **20**, 1259 – 1277.

Díaz-García, J.A. and Leiva, V. (2005). A new family of life distributions based on the elliptically contoured distributions. *Journal of Statistical Planning and Inference*, **128**, 445 – 457.

Stasinopoulos, D.M. and Rigby, R.A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, **23**, 1 – 10.

Linear mixed models innovative methodological approaches in pseudo-panels

Vicente Núñez-Antón[1], Ainhoa Oguiza[1], Jorge Virto[1]

[1] Departamento de Econometría y Estadística (E.A. III), Universidad del País Vasco UPV/EHU, Avenida Lehendakari Aguirre, 83, E-48015 Bilbao, Spain.

E-mail for correspondence: ainhoa.oguiza@ehu.eus

Abstract: In this paper we discuss a model for pseudo-panel data when some but not all of the individuals stay in the sample for more than one time-period and, thus, we propose to model the individuals' time dependency by using a linear mixed effects model with the use of R programming. Data correspond to the Basque labor market for the period 2005-2012, which includes an economical crisis or recession period. Results suggest that the effect of the economical crisis on employment rates is not homogeneous or similar for males and females, as well as for the individuals' different educational levels.

Keywords: Economical crisis; Educational level; Employment rates; Gender; Women's literacy; Linear mixed models; Pseudo-panel data.

1 Introduction

When we have observations on a set of individuals along different periods of time, we say that we have a "panel of data". However, we may have observations on sets of individuals that change from one period to another, which do not constitute a panel of data. An example of this are the data obtained at the Family Expenditure Surveys which are held by many countries.

The main difference in our approach with respect to previous research is that we do not consider the case of independent samples, but rather introduce time dependence between them. Thus, while using pseudo-panels, the work presented in this paper follows a different approach. We model this correlation structure time dependence by using a linear mixed effects model methodological approach, as is usually done in longitudinal or panel data analysis. Linear mixed effects models are considered as one of the more robust methods in Statistics, and have become a really appealing methodology for the analysis of panel data.

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

2 Labor market data

The data we use come from a large data base (i.e., the PRA) obtained from EUSTAT, The Basque Institute of Statistics, (1986). The database includes 40 quarterly periods (i.e., ten years), going from the first quarter in the year 2005 (i.e., 2005-I) to the las quarter in the year 2012 (i.e., 2012-IV). One of the most interesting characteristics about this database is that it is based on a continuous probabilistic sample; that is, it includes a panel of households that is continuously changing or, more specifically, continuously being updated. The PRA current sample has an approximate sample size of about 5000 households per quarter (with an approximate total of about 13500 individuals), and a so-called rotation or household update of one eighth from one to the next quarter. That is, the same household remains in the sample for eight quarterly periods (i.e., two years) and, then, it is replaced or is no longer in the sample (EUSTAT, 1986).
The variables selected for the analysis are: employment rate, which will be the response variable, and gender, educational level and age, which will be used as explanatory variables.

3 Methodological issues

3.1 Construction of the data cohorts

As we have already mentioned, and given that individuals in the data under study do not remain in the sample for the whole period, the first step in the analysis consists on the specific proposal to build the so-called "representative individuals" or "standard individuals" (i.e., individual's cohorts) that we are able to follow along time for the complete period under study. In this way, we build the so-called pseudo panel data of individuals (Deaton, 1985). In our specific data set, standard individuals or cohorts were constructed as a function of the variables gender, educational level and age. As for gender, we have males an females; as for educational level, we have three different categories: primary school, high school and university studies. Finally, as for age, the data base provides this variable already divided by the five age intervals listed here: 16-24, 25-34, 35-44, 45-54 and 55-65 years old. Therefore, we should have a total of 30 different standard individuals or cohorts for the possible level combinations of gender, educational level and age. However, there are not enough individuals having only primary school studies for the first three age intervals, leading, thus, to having at the end only 24 possible standard individuals or cohorts to be analyzed in the final data set. In this way, as it is usually done when dealing with pseudo panel data settings, we have averaged the employment rates for the different individuals belonging to each of these standard individual categories, so that we observe the same cohort or standard individual over time. Moreover, and given that from their own construction, observations corresponding to standard individuals or cohorts are not independent over time, this characteristic or behavior needs to be included in any methodological approach considered for their analysis. Furthermore, as individuals used to compute the corresponding employment rate averages for the standard individuals or cohorts remain in the sample for several periods of time, a clear time dependence is present and should be appropriately modelled.

3.2 The Model

A natural, well accepted and known methodological way of modelling this correlation structure time dependence is by using a linear mixed effects model methodological approach, as is usually done in longitudinal or panel data analysis. These models can be seen as an extension of the classic regression model for cross sectional data when random effects are included to take into account the possible existing heteroscedasticity for the different individuals.

In our model proposal, for the fixed effects in the linear mixed effects models setting, we consider a global quadratic time trend (i.e., it includes the terms on t and t^2), as well as the corresponding effects for the variables educational level, gender and age. For the variable educational level, we have used three dummy variables, $Stud(1)$, $Stud(2)$ and $Stud(3)$, corresponding to the primary school, high school and university studies educational levels, respectively. The category university studies will be used as reference level and, thus, it will not be included in the model. For the variable gender, we have used two dummy variables, one for males (Mal) and one for females (Fem). The former will be used as reference level and, thus, it will not be included in the model. Finally, for the variable age, we have used five dummy variables, $Age(1)$, ..., $Age(5)$, corresponding to the age intervals 16-24, 25-34, 35-44, 45-54 and 55-65 years old, respectively. The age interval 16-24 will be used as reference level and, thus, it will not be included in the model. As for the random effects, we propose the use of an individual-specific quadratic time trend that would allow for individual differences with respect to the global trend given by the fixed effects global quadratic time trend. Therefore, the proposed linear mixed effects model for the response variable given by the employment rate for the standard individual or cohort i at time j, O_{ij}, will be given by:

$$\begin{aligned}
O_{ij} \quad = \quad & \overbrace{\beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \sum_{k=1}^{k=2} \delta_k Stud(k)_{ij} + \lambda Fem_{ij} + \sum_{k=2}^{k=5} \gamma_k Age(k)_{ij} + \sum_{k=2}^{k=5} \theta_k \left(Fem_{ij} * Age(k)_{ij}\right) +}^{Fixed\ Effects} \\
& \overbrace{+ b_{0i} + b_{1i} t_{ij} + b_{2i} t_{ij}^2}^{Random\ Effects} + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2), \quad i = 1, 2, \ldots, 24; \quad j = 0, 1, 2, \ldots, 7 \\
& (b_{0i}, b_{1i}, b_{2i})^T \sim N(0, D), \quad D = \begin{pmatrix} \sigma_{b_0}^2 & \sigma_{b_0 b_1} & \sigma_{b_0 b_2} \\ \sigma_{b_0 b_1} & \sigma_{b_1}^2 & \sigma_{b_1 b_2} \\ \sigma_{b_0 b_2} & \sigma_{b_1 b_2} & \sigma_{b_2}^2 \end{pmatrix} \qquad (1)
\end{aligned}$$

4 Model Estimation

TABLE 1. Estimates and standard deviations for linear mixed effects models parameters. Restricted maximum log-likelihood function and Akaike's information criterion (AIC) values are included.

	Parameter	Fixed Effects	Mixed Effects
REML Log-Lik value		201.353	366.016
AIC		-474.267	-690.031
Shapiro-Wilk's normality p-value		0.0004	0.7036
Fixed effects			
Intercept	β_0	0.307 (0.021)	0.252 (0.047)
Linear Time	β_1	0.006 (0.008)	0.006 (0.003)
Quadratic Time	β_2	-0.002 (0.001)	-0.002 (0.001)
Primary School	δ_1	-0.241 (0.016)	-0.321 (0.034)
High School	δ_2	-0.094 (0.011)	-0.143 (0.023)
Female	λ	0.009 (0.024)	0.051 (0.052)
Age Group 2: 25-34 years	γ_2	0.555 (0.024)	0.548 (0.052)
Age Group 3: 35-44 years	γ_3	0.656 (0.024)	0.712 (0.052)
Age Group 4: 45-54 years	γ_4	0.688 (0.022)	0.719 (0.049)
Age Group 5: 55-65 years	γ_5	0.415 (0.022)	0.533 (0.049)
Female & Age Group 2	θ_2	-0.044 (0.034)	-0.023 (0.074)
Female & Age Group 3	θ_3	-0.156 (0.034)	-0.121 (0.074)
Female & Age Group 4	θ_4	-0.227 (0.031)	-0.141 (0.068)
Female & Age Group 5	θ_5	-0.206 (0.031)	-0.151 (0.068)
Random effects			
Variance of intercepts	$\sigma^2_{b_0}$	–	0.0097
Variance of time effects	$\sigma^2_{b_1}$	–	0.2078
Variance of time square effects	$\sigma^2_{b_2}$	–	0.0169
Covariance parameters for D matrix			
Covariance of intercepts, time	σ_{b_0,b_1}	–	-0.0389
Covariance of intercepts, time square	σ_{b_0,b_2}	–	-0.0071
Covariance of time, time square	σ_{b_1,b_2}	–	0.0393
Variance parameter for the error term			
Residual variance	σ^2	0.0046	0.0003

Approximate standard deviations for the fixed-effects estimates are included in parentheses.

4.1 Alternative specifications

In the modelling proposal, we have considered two alternative model specifications: a fixed effects model, where $\sigma^2_{b_0} = \sigma^2_{b_1} = \sigma^2_{b_2} = \sigma_{b_0b_1} = \sigma_{b_0b_2} = \sigma_{b_1b_2} = 0$; and the more general mixed effects model given by (1). Both model proposals were estimated with restricted maximum likelihood estimation methods (REML) with the use of the function lme (Pinheiro et al., 2014) implemented in the R statistical software program (R Core Team, 2014). Table 1 includes a summary of the result obtained when fitting both model proposals, including estimates, standard deviations and goodness-of-fit measures, such as the restricted maximum log-likelihood function and the Akaike's information criterion values. In addition, the Shapiro-Wilk's normality test for models' appropriateness was also conducted and their p-values are also reported.

Model selection between the fixed and random effects models we have fitted was based on the likelihood ratio test (LRT) and its restricted log-likelihood function value. The corresponding REML-based LRT value equals 329.32 (p-value < 0.0001), which clearly rejects the fixed effects model at the usual $\alpha = 0.05$ significance level. We have also considered several intermediate alternative models to the more general random effects model in (1). However, the corresponding REML-based LRT concluded that the the aforementioned proposed linear mixed effects model is the best fitting model. Among those intermediate models, we have estimated a model with diagonal D matrix i.e., ($\sigma_{b_0b_1} = \sigma_{b_0b_2} = \sigma_{b_1b_2} = 0$), and compared it with the general model in (1). The corresponding REML-based LRT clearly rejected this reduced model (p-value$= 0.0008$).

As recommended in the literature, model selection between the different possible fixed effects models we have fitted was based on the likelihood ratio test and its log-likelihood function value (ML-based LRT). Finally, we now proceed to interpret the results obtained in this model and reported in Table 1. For the sake of brevity of exposition and the difficulty of including all possible test results therein, p-values are not reported in Table 1.

4.2 Interpretation and conclusions

First of all, the clear and significant lineal and quadratic trends (regression coefficients β_1 and β_2) explain and model the employment rate evolution and behavior along time. We can observe that, during the pre-crisis period, corresponding to the years 2005 to 2008, employment rates remain basically constant, and then, in the year 2009, they start a clear descending trend.

With regard to the effect the variable educational level has on employment rates, we can see a clear positive effect of this variable on the response variable under study. Individuals with university studies have a mean estimated employment rate about 32.1% above the one for individuals with only primary school studies (estimated regression coefficient δ_1), and about 14.3%

above the one for individuals with only high school studies (estimated regression coefficient δ_2). Therefore, these results support the claim that a higher educational level significantly increases employment rates.
In addition, results also indicate that there are clear employment rate differences for the different age intervals (estimated regression coefficients $\gamma_2, \gamma_3, \gamma_4$ and γ_5) for males (the reference gender level), and with respect to the reference age interval given by individuals in the age interval 16-24 years old. In this way, male individuals in the interval 35-54 years old have a mean estimated employment rate more than 70% above the one for individuals in the reference age interval 16-24 years old, whereas this difference becomes smaller for individuals in the age intervals 25-34 and 55-65 years old, being close to 54% for these two cases. In addition, the proposed model also shows the clear existing difference between male and female employment rates, which, as expected, does somehow depend on the individuals' age through the interaction terms in the model (estimated regression coefficients $\theta_2, \theta_3, \theta_4$ and θ_5). On the one hand, the smallest employment rate difference between males and females corresponds to younger individuals, in the age interval 25-34 years old, a difference only about 3% higher for female individuals (estimated regression coefficient $\lambda + \theta_2$). On the other hand, the largest employment rate difference between males and females corresponds to individuals in the age intervals 45-54 and 55-65 years old, about 10% higher for male individuals (estimated regression coefficient $\lambda + \theta_5$).

Acknowledgments: Special thanks to the Instituto Vasco de Estadística (EUSTAT). This work was supported by Ministerio de Economa y Competitividad, FEDER, the Department of Education of the Basque Government (UPV/EHU Econometrics Research Group), and Universidad del Pas Vasco UPV/EHU under research grants MTM2013-40941-P, ECO2013-40935-P, IT-642-13 and UFI11/03.

References

Deaton, A. (1985). Panel data from time series of cross-sections. *Journal of Econometrics* **30**, 109-126.

EUSTAT (1986). Encuesta Continua de la Población en Relación con la Actividad. *Eusko Jaurlaritza/Gobierno Vasco: Vitoria/Gasteiz.*

Oguiza, A., Gallastegui, I., and Núñez-Antón, V . (2012). Analysis of pseudo-panel data with dependent samples. *Journal of Applied Statistics* 39(9), 1921-1937.

Pinheiro, J.C., Bates, D.M., DebRoy, S., and Sarkar, D. (2014). nlme: Linear and Nonlinear Mixed Effects Models. *R Core Team. R package version 3.* 1-118.

CTP distribution versus other usual models for count data

M. J. Olmo-Jiménez[1], J. Rodríguez-Avi[1], V. Cueva-López[1]

[1] University of Jaén, Spain

E-mail for correspondence: vcl00006@red.ujaen.es

Abstract: The Complex Triparametric Pearson (CTP) distribution is a count data model useful for modelling situations with under- and overdispersion. The aim of this work is to compare the CTP distribution with the negative binomial, the complex biparametric Pearson and the univariate generalized Waring distributions, all of them for overdispersed count data, the latter with three parameters. The comparison is made through the probability mass function, the skewness and kurtosis coefficients and the Kullback-Leibler divergence. Finally, some examples in the socio-economic field are included to illustrate the versatility of the CTP distribution versus the aforementioned distributions.

Keywords: Overdispersion; Underdispersion; Models for count data.

1 The CTP distribution

The Complex Triparametric Pearson (CTP) distribution, with parameters $a, b \in \mathbb{R}$, and $\gamma > \max(0, 2a)$, was developed by Rodríguez-Avi *et al.* (2004). It is a count-data distribution of infinite range generated by the Gauss hypergeometric function ${}_2F_1(a+ib, a-ib; \gamma; 1)$, where i is the imaginary unit. Its probability mass function (pmf) is given by:

$$f(x) = f_0 \frac{(a+ib)_x (a-ib)_x}{(\gamma)_x} \frac{1}{x!}, \quad x = 0, 1, \ldots$$

where $(\alpha)_r = \Gamma(\alpha + r)/\Gamma(\alpha), \alpha > 0$ and $f_0 = \frac{\Gamma(\gamma-a-ib)\Gamma(\gamma-a+ib)}{\Gamma(\gamma)\Gamma(\gamma-2a)}$ is the normalizing constant.

This distribution is a generalization of the Complex Biparametric Pearson (CBP) distribution (Rodríguez-Avi *et al.*, 2003), since the latter appears when $a = 0$.

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

The main properties of the *CTP* distribution are summarized as follows (for an exhaustive review of these properties see Rodríguez-Avi *et al.*, 2004):

1. There are explicit expressions for the mean and the variance in terms of the parameters of the model, that is,

$$\mu = \frac{a^2+b^2}{\gamma-2a-1}, \quad \sigma^2 = \mu\frac{\mu+\gamma-1}{\gamma-2a-2}.$$

   To guarantee the existence of the mean and the variance it is clear that $\gamma > 2a+1$ and $\gamma > 2a+2$, respectively.

2. If $\frac{(a-1)^2+b^2}{\gamma-2a+1} \in \mathbb{Z}$, the distribution has two consecutive modes in this value and the previous one. In other case, the distribution is unimodal with mode in the integer part of that value. As a consequence, the pmf is $J-$shaped or bell-shaped. Moreover, it is a right skewed distribution.

3. It is underdispersed if $a < -(\mu+1)/2$, equidispersed if $a = -(\mu+1)/2$ or overdispersed if $a > -(\mu+1)/2$. In particular, if $a \geq 0$ the *CTP* is always overdispersed. This property makes the *CTP* distribution more versatile to model a dataset.

4. A sufficient condition to be infinitely divisible (i.d.) is that $a > -0.5$ and $\gamma > (a^2+b^2)/(1+2a)$. So, if $a < -0.5$ the *CTP* distribution is not i.d.

5. It converges to the Poisson distribution when γ and $a^2+b^2 \to \infty$ with the same order of convergence and to the normal distribution when γ and $\sqrt{a^2+b^2}$ have the same order of convergence.

2 Comparison with other count data distributions

To understand the differences between the *CTP* distribution and other common distributions for count data, we compare them through the probability mass function and the Kullback-Leibler divergence. Specifically, we consider in the comparison three distributions wich cope with overdispersed count data: negative binomial (*NB*), *CBP*, and univariate generalized Waring (*UGW*) (see Johnson *et al.*, 2005), the latter with three parameters like the *CTP*.

Since it makes no sense to compare two arbitrary distributions, we fix the first two moments: mean (μ) and variance (σ^2). The pmfs for the NB, CBP, UGW and *CTP* distributions for several values of μ and σ^2 appear in Figure 1. It is not possible to fix the third moment to compare the *CTP* and the *UGW*, both with three parameters, because the first one reduces to a particular case of the second (both are generated by the Gauss hypergeometric function with $\lambda = 1$). So, we consider $b = 1$ in the CTP distribution and $k = 2, 10$ in the UGW. It can be observed that there are differences among these distributions which are more evident as μ and σ^2 increase.

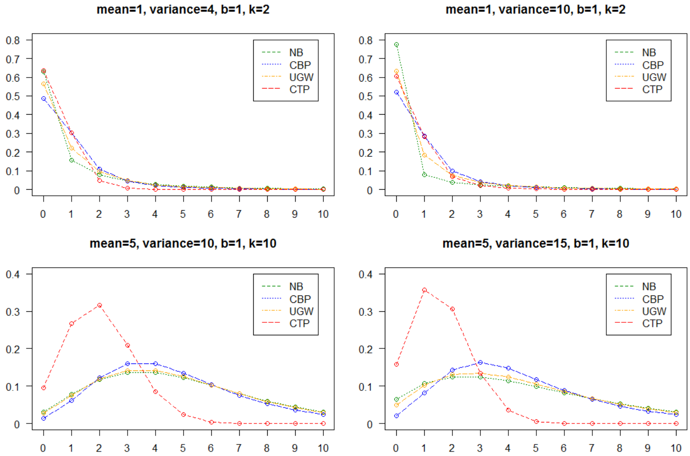


FIGURE 1. P.m.f. of the NB, CBP, UGW and CTP distributions for several values of the mean and the variance.

In addition, Figure 2 shows the values of the KL divergence between the CTP and NB, CBP and UGW distributions, respectively (and vice versa) in terms of σ^2. For the sake of brevity we only consider $\mu = 4$, $\gamma = 4$ in the CTP distribution and $k = 9$ in the UGW. We observe that the CTP is closer to the CBP than the UGW or the NB and the differences increase as σ^2 increases.

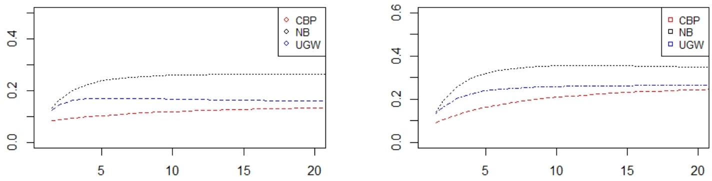


FIGURE 2. Kullback Leibler divergence between the CTP and the NB, CBP and UGW distributions (and vice versa) for several values of the variance.

3 Application examples

We include two examples in the educational field to illustrate that the CTP distribution can provide more accurate fits than other usual distributions for count data (NB, CBP and UGW). Specifically, we consider the number of nursery and primary schools by municipality in Andalusia in 2011. Table 1 contains the values of the Akaike information criterion (AIC) related to the maximum likelihood (ML) fits. In both cases the best fit is that corresponding to the CTP model.

TABLE 1. AIC of the ML fits for data about the number of nursery and primary schools by municipality in Andalusia in 2011.

	NB	CBP	UGW	CTP
Nursery schools	3679.745	3369.261	3479.611	**3355.704**
Primary schools	3280.572	2892.708	3068.55	**2722.265**

4 Conclusions

The comparison procedure shows that the CTP distribution has a differential shape with respect to the most common count data distributions. An important difference with the NB, CBP and UGW distributions is that the CTP is useful for both under- and overdispersed data. Also, it can model real data more accurately when the modal value is not 0 and with low overdispersion. The shape and the Kullback-Leibler divergence also reveal important differences.
In order to use the CTP as a model for real data, we compare the fits obtained for variables related to educational infrastructures in Andalusia (Spain). Results show that the CTP is an adequate model for these type of data.

References

Johnson, N.L., Kemp, A.W. and Kotz, S. (2005). *Univariate Discrete Distributions*. John Wiley & Sons, Inc., 3rd Edition.

Rodríguez-Avi, J., Conde-Sánchez, A. and Sáez-Castillo, A.J. (2003). A new class of discrete distribution with complex parameters. *Statistical Papers*, **44**(1), 67 − 88.

Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A.J. and Olmo-Jiménez, M.J. (2004). A triparametric discrete distribution with complex parameters. *Statistical Papers*, **45**(1), 81 − 95.

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Semiparametric log-symmetric regression models for censored data

Gilberto A. Paula[1] and Luis Hernando Vanegas[2]

[1] Instituto de Matemática e Estatística, Universidade de São Paulo, Brazil
[2] Departamento de Estadística, Universidad Nacional de Colombia, Colombia

E-mail for correspondence: `giapaula@ime.usp.br, havanegasp@gamil.com`

Abstract: In this paper we propose a class of semiparametric accelerated failure time models in which the failure times follow a specific log-symmetric distribution and non-informative left- or right-censoring may be considered. A reweighed iterative process based on the back-fitting algorithm is derived for the estimation of the maximum penalized likelihood estimates. Model selection procedures are proposed and a real data set is analyzed in `R` under the proposed models.

Keywords: Accelerated failure time models; Asymmetric distributions; Non-informative censoring; Robust estimation.

1 Introduction

The aim of this paper is to propose a class of semiparametric accelerated failure time models under the presence of non-informative left- or right-censored data in which the failure times follow a specific log-symmetric distribution. The log-symmetric class (see, for instance, Vanegas and Paula, 2016) contains various asymmetric continuous distributions with lighter or heavier tails than the log-normal one, such as log-Student-t, Birnbaum-Saunders, log-power-exponential, harmonic law and log-slash, among others. The location and scale parameters are modelled in a semiparametric way with the nonparametric components being approximated by B-splines or P-splines. From an appropriate penalized log-likelihood function a back-fitting algorithm is derived for the parameter estimation. Some model selection procedures are also derived and a real data set is analyzed in `R` under the proposed models.

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

2 The model

We will consider the following accelerated failure time model:

$$T_i = \eta_i \epsilon_i^{\sqrt{\phi_i}}, \tag{1}$$

for $i = 1, \ldots, n$, where T_i denotes the failure (or censoring) time, η_i is the location parameter (median) and $\phi_i > 0$ is the scale parameter (skewness) of the ith experimental unit, with $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{LS}(1, 1, g(\cdot))$ (standard log-symmetric distribution), where $g(u) > 0$ denotes the density generator for $u > 0$ and $\int_0^\infty u^{-\frac{1}{2}} g(u) du = 1$. Each log-symmetric distribution may include extra parameters, ζ_1 and ζ_2, which are estimated separately by a goodness-of-fit measure. The survival function of T_i may be expressed as $S_{T_i}(t) = S_{\epsilon_i}\{(t/\eta_i)^{1/\sqrt{\phi_i}}\}$, where $S_{\epsilon_i}(\cdot)$ denotes the survival function of ϵ_i. Moreover, the median and skewness parameters are modelled as

$$\eta_i = \exp\{\mathbf{x}_i^\top \boldsymbol{\beta} + \sum_{j=1}^{p'} f_{\eta_j}(a_{ij})\} \quad \text{and} \quad \phi_i = \exp\{\mathbf{z}_i^\top \boldsymbol{\gamma} + \sum_{k=1}^{q'} f_{\phi_k}(b_{ik})\}, \tag{2}$$

for $i = 1, \ldots, n$, where $\mathbf{x}_i^\top = (x_{i1}, \ldots, x_{ip})$ and $\mathbf{z}_i^\top = (z_{i1}, \ldots, z_{iq})$ contain explanatory variable values as well as $(a_{i1}, \ldots, a_{ip'})^\top$ and $(b_{i1}, \ldots, b_{iq'})^\top$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^\top$ are the regression coefficients and $f_{\eta_j}(\cdot), j = 1, \ldots, p'$, and $f_{\phi_k}(\cdot), k = 1, \ldots, q'$, are continuous, smooth and unknown functions, which are approximated by B-splines or P-splines.

3 Parameter estimation

Applying the Gauss-Seidel method (see, for instance, Hastie and Tibshirani, 1990), the $(u+1)$th step of the iterative process for obtaining the maximum penalized likelihood estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\tau}_{\eta_1}, \ldots, \boldsymbol{\tau}_{\eta_{p'}}$ by fixing $\boldsymbol{\gamma}$ and $\boldsymbol{\tau}_{\phi_1}, \ldots, \boldsymbol{\tau}_{\phi_{q'}}$ may be expressed as

$$\begin{aligned}
\boldsymbol{\beta}^{(u+1)} &= \{\mathbf{X}^\top \mathbf{D}_\eta^{(u)} \mathbf{X}\}^{-1} \mathbf{X}^\top \mathbf{D}_\eta^{(u)} \{\tilde{\mathbf{y}}^{(u)} - \sum_{\ell \neq 0} \mathbf{N}_{\eta_\ell} \boldsymbol{\tau}_{\eta_\ell}^{(u+1)}\} \\
\boldsymbol{\tau}_{\eta_j}^{(u+1)} &= \{\mathbf{N}_{\eta_j}^\top \mathbf{D}_\eta^{(u)} \mathbf{N}_{\eta_j} + \lambda_{\eta_j} \mathbf{M}_{\eta_j}\}^{-1} \mathbf{N}_{\eta_j}^\top \mathbf{D}_\eta^{(u)} \{\tilde{\mathbf{y}}^{(u)} - \sum_{\ell \neq j} \mathbf{N}_{\eta_\ell} \boldsymbol{\tau}_{\eta_\ell}^{(u+1)}\}
\end{aligned}$$

for $j = 1, \ldots, p'$ and $u = 0, 1, \ldots$. The iterative process above should be alternated with the following iterative process:

$$\begin{aligned}
\boldsymbol{\gamma}^{(s+1)} &= \{\mathbf{Z}^\top \mathbf{D}_\phi^{(s)} \mathbf{Z}\}^{-1} \mathbf{Z}^\top \mathbf{D}_\phi^{(s)} \{\tilde{\mathbf{z}}^{(s)} - \sum_{\ell \neq 0} \mathbf{N}_{\phi_\ell} \boldsymbol{\tau}_{\phi_\ell}^{(s+1)}\} \\
\boldsymbol{\tau}_{\phi_k}^{(s+1)} &= \{\mathbf{N}_{\phi_k}^\top \mathbf{D}_\phi^{(s)} \mathbf{N}_{\phi_k} + \lambda_{\phi_k} \mathbf{M}_{\phi_k}\}^{-1} \mathbf{N}_{\phi_k}^\top \mathbf{D}_\phi^{(s)} \{\tilde{\mathbf{z}}^{(s)} - \sum_{\ell \neq k} \mathbf{N}_{\phi_\ell} \boldsymbol{\tau}_{\phi_\ell}^{(s+1)}\}
\end{aligned}$$

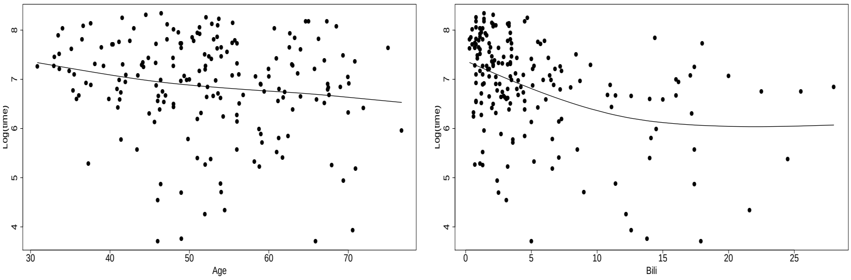


FIGURE 1. Dispersion graphs between log(time) and age (left) and between log(time) and bili (right) for the uncensored data from the PBC data set.

for $k = 1, \ldots, q'$ and $s = 0, 1, \ldots$, where $\mathbf{X}$ and $\mathbf{Z}$ are model matrices, $\mathbf{D}_\eta$ and $\mathbf{D}_\phi$ are diagonal matrices that depend on the assumed log-symmetric distribution, $\mathbf{N}_{\eta_j}$, $\mathbf{N}_{\phi_k}$, $\mathbf{M}_{\eta_j}$ and $\mathbf{M}_{\phi_k}$ are matrices related with the spline basis functions, whereas τ_{η_j} and τ_{ϕ_k} are the respective coefficient vectors and λ_{η_j} and λ_{ϕ_k} the smoothing parameters, $j = 1, \ldots, p'$ and $k = 1, \ldots, q'$. In addition, $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{z}}$ are pseudo-response vectors whereas $\mathbf{N}_{\eta_0} = \mathbf{X}$ and $\mathbf{N}_{\phi_0} = \mathbf{Z}$. The smoothing parameters are estimated by the AIC criterion.

4 Application

As illustration we will consider part of the data set available in the object pbc of the R package survival, from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. The data considered consist of $n = 418$ observations and include the response variable, time, that denotes the number of days between registration and the earliest of death, transplantation, or study analysis in July, 1986; and other independent variables, such as status (0: alive at last contact, 1: liver transplant, 2: death), edema (0: no edema, 0.5: edema present without diuretics or edema resolved by diuretics, 1: edema despite diuretic therapy), age, age in years and bili, level of serum bilirubin (mg/dl). By considering the status condition “alive at last contact”as censoring, one has 55.50% of censored data.

From Fig. 1 one may notice a linear tendency between log(time) and age (with some indication of varying dispersion) and a nonlinear tendency between log(time) and bili. Then, we will try to select a suitable model in the class (1)-(2) with systematic components $\log(\eta_i) = \beta_0 + \beta_1 \texttt{edema0.5}_i + \beta_2 \texttt{edema1}_i + \beta_3 \texttt{age}_i + f_\eta(\texttt{bili}_i)$ and $\log(\phi_i) = \gamma_0 + \gamma_1 \texttt{age}_i$, for $1 = 1, \ldots, 418$, where edema0.5 and edema1 denote dummy variables whereas $f_\eta(\texttt{bili}_i)$ is a nonparametric function approximated by a P-spline with $n = 10$ knots. Comparing various log-symmetric models, the best

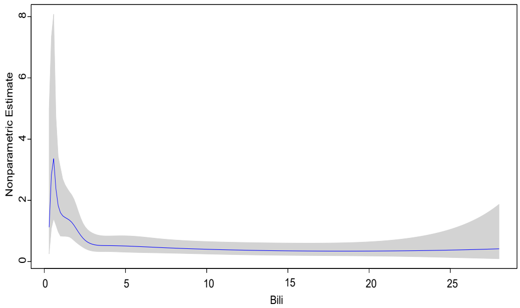


FIGURE 2. Simultaneous 95% confidence intervals for $f_\eta(\texttt{bili})$.

fit (smallest AIC) was attained under `time` $\sim$ log-powerexp(η, ϕ, ζ) with $\hat{\zeta} = 0.55$. The varying dispersion was not significant.

The models were fitted by the function `ssym.l2()` in the `R` package `ssym` (Vanegas and Paula, 2016). We found the parameter estimates (approximated standard errors) $\hat{\beta}_0 = 9.066(0.298)$, $\hat{\beta}_1 = -0.489(0.147)$, $\hat{\beta}_2 = -1.437(0.231)$, $\hat{\beta}_3 = -0.022(0.005)$ and $\log(\hat{\phi}) = -1.186(0.123)$. The smoothing parameter and the effective degrees of freedom were estimated as $\lambda = 0.567$ and $\text{df}(\lambda) = 7.037$, respectively. Fig. 2 presents the simultaneous 95% confidence intervals for the nonparametric function $f_\eta(\texttt{bili})$. The normal probability plot with the quantile residuals from the selected model does not present unusual features.

Therefore, based on the above results, one may conclude that the median failure time (as well as any quantile) decreases as the age increases, is larger for “no edema”condition and presents the behaviour in Fig. 2 according with the level of serum bilirubin.

Acknowledgments: Special Thanks to CNPq and FAPESP, Brazil.

References

Hastie T.J. and Tibshirani R.J. (1990). *Generalized Additive Models.* London: Chapman & Hall.

Vanegas L.H. and Paula G.A. (2016). `ssym`: Fitting Semiparametric Symmetric Regression Models, R package version 1.5.5, http://CRAN.R-project.org/package=ssym.

Vanegas, L.H. and Paula, G.A. (2016). Log-symmetric distributions: statistical properties and parameter estimation. *Brazilian Journal of Probability and Statistics*, **30**, 196-220.

Statistical Modelling of Submissions to Municipalities in the Czech Republic

Anna Pidnebesna[1], Kateřina Helisová[1], Jiří Dvořák[2], Radka Lechnerová[3], Tomáš Lechner[4]

[1] Czech Technical University in Prague, Czech Republic
[2] Charles University in Prague, Czech Republic
[3] Private College of Economic Studies in Prague, Czech Republic
[4] University of Economics in Prague, Czech Republic

E-mail for correspondence: pidneann@fel.cvut.cz

Abstract: During last years, space-time point processes became a usefull tool in modelling different random events from many fields such as medicine, biology or economy. This work concerns using these processes for modelling of submissions to municipalities in the Czech Republic. The positions and times of submissions form the space-time point pattern which is to be analysed. The most appropriate model is chosen from the list of continuous and discrete methods of modelling. The suitability of this model is justified through classical methods of spatial statistics using simulation studies. This work is extract of [Pidnebesna et al. (2016)].

Keywords: Cluster process; Empirical distribution; Statistical analysis; Statistical modelling; Submissions to municipalities.

1 Introduction

Modelling of space-time point patterns has become widely used in the recent time. Our purpose is to use the existing theory and technical availability for analysing and modelling the real-life processes.

We have a large amount of data describing submissions to municipalities in the Czech Republic. The data is observed over a long period of time and form an inhomogeneous space-time point pattern consisting of thousands of points.

From the nature of the data it follows that using of space-time modelling methods can give an suitable result in modelling such type of processes. The theory of spatio-temporal processes is a new field which is still under development, but there already exist many useful results. In the recent

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

papers, there were developed second-order characteristics for the inhomogeneous spatio-temporal point patterns, the separability properties were studied using mentioned characteristics and methods, different types of inhomogeneity were described, and the procedure of the estimate of wide class of inhomogeneous processes were provided.

A significant part of the conducted analysis in our work is based on the research mentioned above. The introduced methods assumes working with the processes in continuous spaces, while in our case the data are roughly discretised both in time and space. In order to provide as precise analysis as possible, we used the continuous approach as well as the discrete one. In this paper we describe the most suitable one.

2 Data description

The data were examined by stemming from the electronic records management systems kept by the municipalities. The data consists of dates, applicants' addresses, agenda and types of communication (electronic, post, personal etc.). However, they were anonymised by the provider so that we have no addresses but only postcodes (ZIP codes) at our disposal. Therefore we identify the communicating person by position of appropriate post office.

We have series containing the date of incoming communication and spatial identification within the territory of the Czech Republic. The aim is to use this data to analyse the spatial behavior and evolution over time. In accordance with this aim we randomly chose a municipality. It is a municipality located about 50km from Prague in the north-west direction in a village having about 2.8 thousand of inhabitants.

The dataset includes 6205 space-time events corresponding to individual submissions. The data were recorded in the time interval from 27th October 2009 to 20th April 2011. During this time interval 370 workdays took place. The submissions came from 214 different ZIP codes.

3 Data model

At the beginning of our research, we proposed a flexible model for inhomogeneous cluster process in the continuous domain. Based on the exploratory analysis we made, it was shown that the data forms cluster process, however the size of the clusters is too small (the scale is less then 1 km in space and less then 1 day in time). Thus it was not possible to infer the precise scale of the clusters from the data and the continuous domain modelling is not appropriate for this dataset. Therefore we started to work with the data in discrete domain.

We tested the hypothesis about independence of spatial and temporal coordinates of the process, which was rejected. Further, we tested the hypothesis that number of repetitions for each space-time point could be described by the Poisson distribution. This hypothesis also was rejected. That is why we focused on the approach based on the empirical distributions.

In order to analyse temporal projection we constructed the periodogram. From it and also from the nature of the data we found that the most important time periods in the data are one month and one week. Further we observed different behavior in different working days caused probably by the fact that the office hours of municipalities are usually on Mondays and Wednesdays. Therefore we divided the days into two groups

$$gr\ A = \{\textit{Monday, Wednesday}\}, \quad gr\ B = \{\textit{Tuesday, Thursday, Friday}\}.$$

Let $t \in \{0, 1, \ldots, T\}$ and consider functions $m(t)$ and $d(t)$ describing to which month and type of (working) day the time t correspond, namely $m(t) = 1$ if t is a day in January, ..., $m(t) = 12$ if t is a day in December, and analogously $d(t) = A$ if $t \in grA$, and $d(t) = B$ if $t \in grB$.

Our analysis of the data showed that we cannot work with the space and the time separately, so we still consider the process of points on the lattice. Now, we are interested in the distribution of the number of points in the knots of the lattice. Thus, we consider the process represented as

$$X = \{(\xi_1, t_1, \eta_1), \ldots, (\xi_N, t_N, \eta_N)\},$$

where η_i are independent random variables. Our analysis showed that this variables could not be described by Poisson distribution therefore we focus on the empirical distribution of η. Thus we work with the collection of independent random variables $\eta'(\xi, m(t), d(t))$ where ξ is the spatial position, and $m(t)$ and $d(t)$ are defined above. Hence for each spatial position ξ, 24 empirical distributions (corresponding to 12 months and 2 different types of day) are to be estimated.

For testing goodness of fit we used two approaches. The first one is the construction of empirical envelopes for the statistics $T_1, \ldots, T_5$:

- T_1 – the number of knots of the lattice with exactly one point (so called unique points),
- T_2 – the number of knots of the lattice with more than one point,
- T_3 – the average number of points in a knot conditionally there is at least one point in the knot,
- T_4 – the average number of points in a knot conditionally there are more than one point in the knot,
- T_5 – the total number of points.

The numerical results are presented in the Table 1. The second approach is construction of the confidence intervals for the average number of submissions per month and per (working) day. Our simulations showed that our model is suitable for the data.

TABLE 1. Testing the model based on the empirical distribution.

Statistics	**q2.5%**	**q97.5%**	**Data** X	**Conclusion**
T_1	3027	3204	3141	*Not reject*
T_2	1077	1177	1139	*Not reject*
T_3	1.92	2.02	1.98	*Not reject*
T_4	3.54	3.83	3.69	*Not reject*
T_5	5994	6343	6205	*Not reject*

4 Conclusion

We realised that behavior of the process of submissions to municipalities in the Czech Republic is very complicated in the sense that it cannot be described neither by classical models of point processes such as Poisson process or cluster processes nor by their modifications. Neither using Poisson distribution for modelling the number of points in the same time and spatial coordinates was successful. Therefore, we suggested the procedure based on empirical approach, which allows us to describe the process of submissions to municipalities and make forecasts of future development. Despite the procedure is time-consuming, because it requires separate calculations for each spatial coordinate and each of the 24 combinations of the month and the type of working day, we can evaluate it as suitable because of its accuracy.

Finally note that the used methods were also applied to the data corresponding to three other randomly chosen municipalities and the obtained results were very similar.

References

Pidnebesna A., Helisová K., Dvořák J., Lechnerová R., Lechner T. (2016). Statistical Analysis and Modelling of Submissions to Municipalities in the Czech Republic. *Bulletin České statistické společnosti.* Submitted.

Gabriel E., Diggle P. J. (2009). Second-order analysis of inhomogeneous space-time point process data. *Statistica Neerlandica*, **63(1)**, 43 – 51.

Møller J., Ghorbani M. (2012). Aspects of second-order analysis of structured inhomogeneous space-time point processes. *Statistica Neerlandica*, **66(4)**, 472 – 491.

Wavelet statistical analysis in multitemporal SAR images

Aluísio Pinheiro[1], Abdourrahmane Atto[2], Emmanuel Trouvé[2]

[1] Department of Statistics, University of Campinas, Brazil
[2] LISTIC Laboratory, Université Savoie Mont Blanc, France

E-mail for correspondence: apinheiro.unicamp@gmail.com

Abstract: Satellite images have received intensive attention from the technology and scientific communities since its inception in 1946. One of the most important applications which is routinely performed with satellites is the timewise monitoring. This multitemporal sequence may be used for: sea glaciers movement assessment; deforestation; changes in urban areas; among other applications. Several methods are available in the literature for the change detection in temporal series of images. Threshold shrinkage based on ratio of subsequent images or on the difference between subsequent images are among the most successfull methods. In general the detection methods in SAR images follow a three-step procedure: (1) pre-processing; (2) pixel-by-pixel comparisons; and (3) image thresholding. We propose wavelet statistical tools which can substitute in one step the three aforementioned steps. We show the advantages in a simulation study as well as in a Pol-SAR Alpine glacier example (TERRA-SAR).

Keywords: Wavelet Analysis; Time Series; Satellite Images

1 Introduction

One of the most important applications which is routinely performed with satellites is the timewise monitoring. This multitemporal sequence may be used for: sea glaciers movement assessment (Fily and Rothrock, 1987); deforestation (Almeida-Filho and Shimabukuro, 2000); and changes in urban areas (Gamba et al., 2006; Ban and Yousif, 2012). A particular application of interest for this paper is the remote monitoring via time series analysis of satellite data of regions that are difficult to access. The real data which is analyzed in this work regards the Alpine region that includes Chamonix, the Mont Blanc, and the Argentierre and Mer de Glace glaciers.

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

Suppose that T images at different times are taken by SAR of an image of interest, say $I_1, \ldots, I_T$. Trang-Lê et al. (2015) and Atto et al. (2013) propose the use of matrices of temporal dissimilarities to both detecting the time-points which present the most significant changes, and, for the time-points so selected, a thorough pixel-by-pixel dissimilarity analysis.
A wavelet representation for each image is obtained, called it $\{A, H, V, D\}$, where A, H, V and D represent the vector of wavelet coefficients for the approximation, horizontal, vertical and diagonal subspaces, respectively. Based on a set of fixed distributions, each image is associated with the best fitted density for the detail coefficients $\{H, V, D\}$. Then the Kulback-Leibler distance for each two pair of time images is computed, generating the so-called matrix for change detection. An analogous procedure is performed for the selected images on smaller images, trying to define where any change has happened. This parametric set-up is suboptimal both statistically and computational, the former because there is no guarantee that the collection of distributions is exhaustive and the latter because the maximum likelihood estimation is numerically intensive, specially for the sub images problem. Just as an example, an image of size 2048x2048, which is not exactly very large, will have 64x64 sub images on each the maximum likelihood must be performed.
What we do here is to substitute the parametric idea by a nonparametric one. First, we do not define beforehand which distributions will be compared. We simply use wavelet density procedures to estimate the densities for each time. Pinheiro and Vidakovic (1997) proposes a non-negative density wavelet estimator. The idea is simple and powerful. When estimating a function g one must represent g in a multiscale analysis, which is only possible only if g is square integrable. For a density f if one estimates $g = \sqrt{f}$, g is always square integrable. Moreover, since $\hat{f}(x) = (\widehat{\sqrt{f}}(x))^2$ one does not get the unsettling (but common in nonparametrics) result of negative density estimates. Moreover, since f integrates to 1 one can smpily renormalize the coefficients, by Parseval's equality, to preserve the comparability of density estimates across time points. The times are then compared by the Hellinger distances between their densities of coefficients.

2 Results

Some important information regarding the wavelet procedures are the following:

1. Due to speckle, soft threshoiding is recommended (Donoho, 1995). The method is robust in the sense that thresholding levels and paradigm do not qualitatively change the dissimilarities matrix, i.e., data points which are similar in one thresholding policy maintains this status for the rest.

2. Compactly supported wavelets (symm8) are used because of its smoothness and filter size.

3. The densities were computed for each detail subspace separately. We should understand that the parametric set of densities for each image (V, H, D) would not necessarily fit to the same density: this was simply numerically too much. Results here show that this is important to increase power.

4. Wavelet thresholding deals with noise reduction and change-point detection at the same time.

There are 11 time points in the TERRA-SAR images which have three different polarizations. The matrices for each polarization are shown as the columns in Figure 1(a). This figure shows the nine (three subspaces V, H and T and three channels), and one can see that the results are consistent for the wavelet subspaces but differ slightly for the channels, as expected.

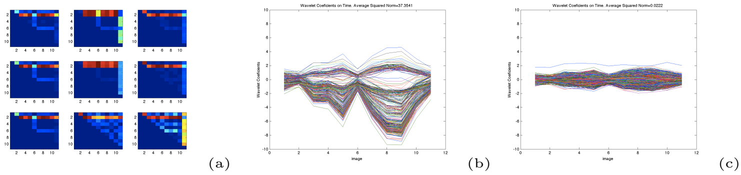


FIGURE 1. (a) Dissimilarity Matrices (b) Relevant wavelet coefficients (c) Unimportant wavelet coefficients

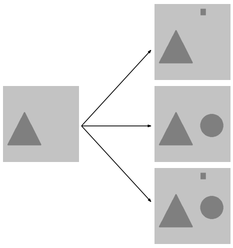


FIGURE 2. Image changes along time

The matrices for each sub image were also used with similar success. However there is also the changes in wavelet coefficients along time which may be used. Figure 1(b)-(c) show the huge differences for coefficients which are due to temporal changes in the images.

A simulation study was also done. The results show that wavelet methods are more robust to misspecification then the parametric competitors. If the 'model' is right, the wavelet procedure loses but not by the same margin it over performs a poorly specified parametric procedure. Figure 2 shows the different simulated changes in the images from the first to the other time points.

The conclusions are that wavelet methods outperform parametric methods because: (i) of the possibility of reducing steps in the analysis; (ii) of its robustness to model missspecification; (iii) it is computationally much less intensive.

References

Almeida-Filho, R., and Shimabukuro, Y.E. (2000). Detecting areas disturbed by gold mining activities through jers-1 SAR images, Roraima state, Brazilian Amazon. *International Journal of Remote Sensing*, **21** (17): 3357–3362, 2000.

Atto, A., Trouvé, E., Berthomieueu, Y., and Mercier, G. (2013). Multi-Date divergence matrices for the analysis of SAR image time series. *IEEE Transactions on Geoscience and Remote Sensing*, **51**(4), 1922–1938.

Ban, Y. and Yousif, O. (2012). Multitemporal spaceborne SAR data for urban change detection in china. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, **5** (4): 1087–1094, 2012.

Donoho, D.L. (1995). De-noising by soft-thresholding. *Information Theory, IEEE Transactions on*, **41** (3): 613–627, 1995.

Fily, M. and Rothrock, D.A. (1987). Sea ice tracking by nested correlations. *Geoscience and Remote Sensing, IEEE Transactions on*, 5: 570–580.

Pinheiro, A., and Vidakovic=, B. (1997). Estimating the square root of a density via compactly supported wavelets. *Computational Statistics and Data Analysis*, **25** (4), 399–415.

Torres, A.and Frery, A.C (2013). SAR image despeckling algorithms using stochastic distances and nonlocal means. *arXiv preprint arXiv:1308.4338*.

Torres,L. Sant'Anna, S.J.S., Freitas, C.C., and Frery, A.C. (2014). Speckle reduction in polarimetric SAR imagery with stochastic distances and nonlocal means. *Pattern Recognition*, **47** (1): 141–157, 2014.

Trang-Lê, T., Atto, A., Trouvé, E., Solikhin, A., and Pinel, V. (2015). Change detection matrix for multitemporal filtering and change analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, **107**, 64–76.

Modelling the proportion of failed courses and GPA scores for engineering major students

Hildete P. Pinheiro[1], Rafael P. Maia[1], Eufrásio Lima Neto[2], Mariana R. Motta[1]

[1] State University of Campinas, Department of Statistics, Brazil
[2] Federal University of Paraiba, Department of Statistics, Brazil

E-mail for correspondence: `hildete@ime.unicamp.br`

Abstract: We use here zero-one inflated beta models with heteroscedasticity to model the proportion of failed courses in Engineering students at the State University of Campinas, Brazil. We also model the GPA score for those students with a heteroscedastic skew t distribution. The database consists of more than 3000 students with Engineering majors who entered in the University from 2000 to 2005. The entrance exam score (ESS) in each subject, some academic variables and their socioeconomic status are considered as covariates in the models. Diagnostics and residual analysis are performed as well.

Keywords: beta inflated models; skew t distribution; heteroscedasticity; educational data.

1 Introdution

The database consists of more than 3000 students with Engineering majors who entered in the State University of Campinas, Brazil from 2000 until 2005. For each student we have all the grades in the required courses taken in the university as well as the proportion of courses he/she failed during his/her Bachelor's degree. We also have their entrance exam scores - EES in each subject (e.g., Mathematics, Portuguese, Geography, History, Biology, Chemistry and Physics), some academic variables as well as their socioeconomic status, which are considered as covariates in the models.
For modelling the proportion of failed courses, we use a zero and one inflated beta regression model (Ospina and Ferrari, 2010) with heteroscedasticity

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

The GPA (Grade Point Average) scores were standardized within each year and course/major and its model is heteroscedastic with a skew t distribution (Azzalini, 1986).

2 Statistical Models

For the proportion of failed courses a zero-one inflated beta model was used and its distribution is given by

$$p(y;\alpha,\gamma,\mu_1,\phi) = \begin{cases} \alpha(1-\gamma), & \text{if } y=0 \\ (1-\alpha)f(y;\mu_1,\phi), & \text{if } y\in(0,1) \\ \alpha\gamma, & \text{if } y=1 \\ 0, & \text{if } y\notin[0,1] \end{cases} \tag{1}$$

with $f(y;\mu_1,\phi) = \dfrac{\Gamma(\phi)}{\Gamma(\mu_1\phi)\Gamma((1-\mu_1)\phi)} y^{\mu_1\phi-1}(1-y)^{(1-\mu_1)\phi-1},\ y\in(0,1)$.
Note that now $E(Y)=\alpha\gamma+(1-\alpha)\mu_1$ and $Var(Y)=\alpha V_1+(1-\alpha)V_2+\alpha(1-\alpha)(\gamma-\mu_1)^2$, with $V_1=\gamma(1-\gamma)$ and $V_2=\mu_1(1-\mu_1)/(\phi+1)$. For more details see Ospina and Ferrari (2010).
We model μ_1 and $\sigma=1/(\phi+1)$ with logit link and $\nu_1=\alpha(1-\gamma)$ and $\tau_1=\alpha\gamma$ with a log link.
For the GPA scores, a skew t type 1 model (ST1) was used with the distribution given by $F_Y(y\mid\mu_2,\sigma^*,\nu_2,\tau_2)=(2/\sigma^*)f_{Z_1}(z)F_{Z_1}(\nu_1 z)$, for $y\in(-\infty,\infty)$, where $\mu_2\in(-\infty,\infty)$, $\sigma^*>0$, $\nu_2\in(-\infty,\infty)$, $\tau_2>0$, $z=(y-\mu_2)/\sigma^*$ and f_{Z_1} and F_{Z_1} are the pdf and cdf of $Z\sim TF(0,1,\tau_2)$, a t distribution with τ_2 degrees of freedom (treated as continuous parameter) and ν_2 is the skewness parameter. For more details of the ST1 distribution see Azzalini (1986).

3 Application and Results

Looking at the histograms for the GPA scores and for the proportions of failed courses, we noticed that the distribution of the GPA scores for the required courses of the Engineering major students are skewed to the left and for the proportion of failed scores, there is a high frequency of zeros and also some frequency of ones. Therefore, we tried to model the GPA with normal, skewed normal, gamma and skewed t distributions. For the proportion of failed courses we used a zero and one inflated beta model. The *gamlss* package, available in the R-Project, was used to fit the models (Rigby and Stasinopoulos, 2005; Stasinopoulos and Rigby, 2007).
In order to understand better the quantitative variables of the data set, we computed Spearman correlations between the quantitative variables where Y_1 is the proportion of failed courses, Y_2 is the GPA score, X_1 is the EES in Physics, X_2 is the EES in Math, X_3 is the EES in Biology, X_4 is the EES

in Chemistry, X_5 is the EES in Portuguese, X_6 is the EES in Geography, X_7 is the EES in History. We found that the highest correlation is between the proportion of failed courses and the GPA (-0.858), which was already expected, but the correlations between the Entrance Exam Scores (EES's) in all subjects are all very low, with the highest correlations being between Physics and Math (0.371), Physics and Chemistry (0.322), Geography and History (0.385). All the correlations between the ESS's and Y_1 and between the EES's and Y_2 are less than 0.2.

The best model for the mean proportion of failed courses (μ_1) with logit link showed significant effects of year (2000, 2001, 2002, 2003, 2004 and 2005), sex, age (< 17, $18 - 20$ and > 21 years), type of High School (Public or Private), type of engineering major, EES in Physiscs, Biology, Chemistry and Portuguese, as well as the status of graduation (Graduated or Did not graduate), the number of semesters in the university (1 to 8, 9 to 10 and ≥ 11 semesters) and an interaction between the latter two. The model of the proportion of zeros (ν_1) with log link had significant effects for year, sex age, family income, type of engineering major as well as the status of graduation and the number of semesters in the university. The best model for the dispersion parameter (σ) with logit link showed significant effects for age, type of engineering major, status of graduation, the number of semesters in the university and an interaction between the latter two. Note that here the larger the $\hat{\sigma}$ value, the larger is the variance of Y_1.

The course (type of engineering major) codes in the models for μ_1 and ν_1 are: 8 = Agricultural Engineering; 9 = Chemical Engineering (daytime); 10 = Mechanical Engineering; 11 = Electrical Engineering (daytime); 12 = Civil Engineering; 13 = Food Engineering (daytime); 34 = Computational Engineering, 39 = Chemical Engineering (night); 41 = ; 43 = Food Engineering (night); 49 = Automation and Control Engineering.

There is not much difference on the proportion of failed courses (μ_1) among the years, but the proportion of zeros (ν_1) seems to be smaller in 2005, followed by 2004 compared with the other years. Male students have greater proportion of failed courses (μ_1) than Female, but the proportion of zeros (ν_1) is smaller for Male than Female students. The younger the students the fewer courses they fail and, of course, the proportion of zeros is higher for younger students.

The higher the score in the Entrance Exam the less courses they fail. The students with lower income have a bigger proportion of zeros. The lowest proportion of failed courses is of the Automation and Control Engineering students followed by Food Engineering (night) students. The course/major with the higher proportion of failed courses is Mechanical Engineering. When looking at the model for ν_1, the highest proportion of zeros is for the Food Engineering students and the smallest is for Civil Engineering students. There is an interaction effect between the status of graduation and the number of semesters in the university in the model for μ_1 For those who graduated, the lower proportion of failed courses is for those who stayed 9

to 10 semesters in the university, followed by those who stayed at least 11 semesters and then those who stayed 1 to 8 semesters. On the other hand, for those who did not graduate, the lower proportion of failed courses is for those who stayed 1 to 8 semesters, followed by those who stayed at least 11 semesters and then those who stayed 9 to 10 semesters. The proportion of zeros (ν_1) is higher for those who graduated. The proportion of ones (τ_1) is estimated to be $exp(-4.05) = 0.017$ and it is significantly different from zero (p-value < 0.0001). The smallest dispersion (σ) is for those who graduated and stayed 9 to 10 semesters, are at most 21 years of age and are from Agricultural Engineering. The biggest dispersion is for those who did not graduate, are over 21 years old, are from the baseline courses (10, 11, 13, 34, 39, 41, 43, 9) and stayed 1 to 8 semesters at the university.
The best model for the GPA is a heteroscedastic skew t with identity link for the mean (μ_2) and log link for the dispersion (σ^*). For the models for GPA and the dispersion, one can say that the younger the student and the greater his/her EES's, the greater is his/her GPA. Students from Public High Schools (PuHS) and Female have greater GPAs. There is an interaction between the status of graduation and the number os semesters For those who graduated and stayed 9 to 10 semesters, the GPA score is greater than those who drop out or were still active. Also, the more semesters they stayed in the University, the worst is their GPA. The skewness parameter was found to be negative ($\hat{\nu}_2 = -0.32$) and significantly different from zero (p-value=0.018), which makes sense, since the distribution of the GPA is skewed to the left. The model for the dispersion parameter (σ^*) showed that only the status of graduation and the number of semesters was found to be significant, but with an interaction between them. The greater variability was found to be for those who did not graduate and stayed 1 to 8 semesters in the university, which makes sense as these are the students who drop out for various reasons. On the other hand, the smallest variability was found to be for those who graduated in 9 to 10 semesters.

References

Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica*, **46**, 199 – 208.

Ospina, R. and Ferrari, S. L. P. (2010). Inflated beta distributions. *Statistical Papers*, **51**, 111 – 126.

Rigby, R. A. and Stasinopoulos D. M.. (2005). Generalized additive models for location, scale and shape,(with discussion). *Appl. Statist.*, **54(3)**, 507 – 554.

Stasinopoulos, D. M. and Rigby, R.A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, **23 (7)**.

Double-saddlepoint approximation for enumeration of tables for exact tests of Hardy-Weinberg proportions

Xiaoyun Quan[1], James G. Booth[2]

[1] Cornell University, Ithaca, NY, USA

E-mail for correspondence: xq44@cornell.edu

Abstract: Exact tests for Hardy-Weinberg proportions are widely applied in population genetics to determine random mating. The feasibility of exact tests requires information on the number of all possible tables of genotype counts given the observed allele counts. However, complete enumeration can be impractical when the allele counts are large. An approximation method has been proposed by Engels (2009). However multiple alleles and sparse data can render the approximation unreliable. An alternative method is proposed here using a double-saddlepoint approximation results in much better performance.

Keywords: Allele; Exact conditional test; Genotype

1 Formulation of the problem

In population genetics, random mating is a crucial assumption to assess before making any further analysis. This can be done by conducting an exact test of Hardy-Weinberg proportions similar to Fisher's exact test for independence in contingency tables. Feasibility of this test requires knowledge of total number of possible tables of genotype counts that are consistent with the observed allele counts. For a vector of genotype counts at k-allele loci $\boldsymbol{g} = (g_{11}, g_{21}, ..., g_{k1}, g_{22}, ..., g_{kk})$ where g_{ij} is the count of genotype ij (a diploid with alleles i and j), the table of the genotype counts is usually displayed in the form of a lower-triangular matrix

$$G = \begin{bmatrix} g_{11} & & & \\ g_{21} & g_{22} & & \\ \vdots & \vdots & \ddots & \\ g_{k1} & g_{k2} & \cdots & g_{kk} \end{bmatrix}$$

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

from which the count of ith allele can be calculated as

$$m_i = 2g_{ii} + \sum_{i>j} g_{ij} \, .$$

Notice that there are $g = k(k+1)/2$ possible genotypes, and the sample size is given by $n = \sum_{i \geq j} g_{ij}$.
Implementation of the exact test requires the enumeration of all possible genotype matrices G that are consistent with the vector of observed allele counts $\boldsymbol{m}$. However, complete enumeration is not feasible with large counts. Engels (2009) proposed Normal Approximation as a modification of Gail and Mantel's method (1977). However the Normal Approximation method can be unreliable when the sample matrix of genotype counts is sparse. In this paper we investigate the use of an alternative approximation proposed by Zipunnikov, Booth and Yoshida (2009) and show that it is much more accurate than the normal approximation.

2 Normal Approximation versus Double-saddlepoint Apprximation

The two methods share the same basic idea that the number of possible tables given a fixed set of allele counts, $|S_{\boldsymbol{m}}|$, is equal to the product of total number of tables for the given sample size without restrictions on allele counts, $|S|$, and the probability assuming a uniform distribution over all possible genotype matrices that the particular fixed set of allele counts is randomly selected, $\mathbb{P}(\boldsymbol{m})$; that is:

$$|S_{\boldsymbol{m}}| = |S| \cdot \mathbb{P}(\boldsymbol{m})$$

The total count $|S|$ is obtained by considering the genotype distribution as 'stars and bars' problem: the genotypes of the n individuals are seen as 'stars' to be randomly placed between $g-1$ bars, which gives

$$|S| = \binom{n+g-1}{g-1} \, .$$

The difference between normal and double-saddlepoint method lies in the approximation of the probability $\mathbb{P}(\boldsymbol{m})$.
The normal approximation method estimates the probability by approximating the allele count distribution (assuming a uniform distribution over the set of genotype matrices) by a multivariate normal distribution. Engels (2009) uses the probability distribution from 'stars and bars' model to determine the expected value and covariance matrix for the set of allele counts. Then, the method of Gail and Mantel (1977) can be adapted to approximate $\mathbb{P}(\boldsymbol{m})$.

The double-saddlepoint approximation method (Zipunnikov et al. 2009) involves fitting a generalized linear model (GLM) in which the genotype counts g_{ij} are assumed to independent geometric variables so that the conditional distribution of the genotype counts given their sum is uniform over all possible genotype mamtrices if the geometric variables are identically distributed. Specifically, consider the GLM with with parameter vector $\boldsymbol{\lambda}$, the log-likelihood is given by

$$l(\boldsymbol{\lambda}) = \boldsymbol{g}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\lambda} + \sum_{i=1}^{g} \log\left(1 - \exp\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\lambda}\right)\right)$$

where $\boldsymbol{x}_i$ is the ith row of $\boldsymbol{X}$. We consider two forms for the design matrix $\boldsymbol{X}$. In model 1, $\boldsymbol{X}_1 = \boldsymbol{2}_g$, a 2-vector of length g. In model 2, $\boldsymbol{X}_2^T\boldsymbol{g} = \boldsymbol{m}$, the vector of allele counts. Notice that, because each genotype consists of exactly two alleles, each row of $\boldsymbol{X}_2$ sums to 2, and hence model 1 is a special case of model 2.

Using this GLM formulation we can estimate the conditional probability of the allele count vector $\boldsymbol{m}$ given n, by the ratio of two saddlepoint density approximations (Daniels, 1954), as suggested by Zipunnikov et al. (2009). Specifically,

$$\mathbb{P}\left(\boldsymbol{m}\right) \approx \hat{f}\left(\boldsymbol{m}|n\right) = \frac{|2\pi\hat{\boldsymbol{I}}_{\boldsymbol{m}}|^{-1/2}e^{-\hat{l}_{\boldsymbol{m}}}}{|2\pi\hat{\boldsymbol{I}}_n|^{-1/2}e^{-\hat{l}_n}}$$

where $\hat{\boldsymbol{I}}_n$ and $\hat{\boldsymbol{I}}_{\boldsymbol{m}}$ are the Fisher information matrices for the two models 1 and 2 respectively, and $\hat{l}_n$ and $\hat{l}_{\boldsymbol{m}}$ are the corresponding maximized likelihoods.

Zipunnikov et al (2009) also suggested further improvements on the accuracy of double-saddlepoint approximation by adding higher order correction terms. Two types of corrections are considered: additive and exponential. Both corrections typically improve the accuracy of the approximation, with the exponential correction generally being the preferred choice.

3 Results comparison

To compare the performance of the various methods, we consider three sample data sets from Engels (2009) paper, all of which had appeared in earlier literature concerning exact tests for Hardy-Weinberg proportions. The data sets are genotype count matrices A, B and C in Figure 1 of Engels (2009) paper. The performances are listed in Table 1 below, where for each sample the exact numbers of table counts and percentage errors of normal approximation ('Normal'), double-saddlepoint (DS) approximation, DS approximation with additive correction term ('DS additive'), and DS approximation with exponential correction term ('DS exp') are given. It can be seen that if the allele counts are large or the genotype matrix

A

$$\begin{bmatrix} 0 & & & \\ 3 & 1 & & \\ 5 & 18 & 1 & \\ 3 & 7 & 5 & 2 \end{bmatrix}\begin{bmatrix} 11 \\ 30 \\ 30 \\ 19 \end{bmatrix}$$

B

$$\begin{bmatrix} 3 & & & & & & & \\ 4 & 2 & & & & & & \\ 2 & 2 & 2 & & & & & \\ 3 & 3 & 2 & 1 & & & & \\ 0 & 1 & 0 & 0 & 0 & & & \\ 0 & 0 & 0 & 0 & 0 & 1 & & \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & \\ 0 & 0 & 0 & 2 & 1 & 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} 15 \\ 14 \\ 11 \\ 12 \\ 2 \\ 2 \\ 1 \\ 3 \end{bmatrix}$$

C

$$\begin{bmatrix} 2 & & & \\ 12 & 24 & & \\ 30 & 34 & 54 & \\ 22 & 21 & 20 & 10 \end{bmatrix}\begin{bmatrix} 68 \\ 115 \\ 192 \\ 83 \end{bmatrix}$$

D

$$\begin{bmatrix} 1236 & & & & & & & & \\ 120 & 3 & & & & & & & \\ 18 & 0 & 0 & & & & & & \\ 982 & 55 & 7 & 249 & & & & & \\ 32 & 1 & 0 & 12 & 0 & & & & \\ 2582 & 132 & 20 & 1162 & 29 & 1312 & & & \\ 6 & 0 & 0 & 4 & 0 & 4 & 0 & & \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\ 115 & 5 & 2 & 53 & 1 & 149 & 0 & 0 & 4 \end{bmatrix}\begin{bmatrix} 6329 \\ 319 \\ 47 \\ 2773 \\ 75 \\ 6702 \\ 14 \\ 2 \\ 333 \end{bmatrix}$$

FIGURE 1. Sample genotype matrices from Engels (2009) paper.

is sparse, the double-saddlepoint approximation outperforms the normal approximation. Furthermore, adding the higher order correction terms improves the accuracy even more.

TABLE 1. Performances of Normal vs D-S approximation methods.

sample	Exact numeration	Normal	DS	DS additive	DS exp
A	162365	-2.36	+7.76	+2.18	+2.01
B	250552020	+15.97	-9.71	+0.55	+0.09
C	1289931294	+24.61	+1.32	+0.39	+0.38

Notice that for the large sample genotype matrix D, complete enumeration via Mathematica (Engels 2009) is not feasible. The normal approximation, DS approximation, DS with additive correction and DS with exponential correction methods estimate the number of tables as $2\times10^{56}, 2.03\times10^{44}, 1.86\times10^{44}$ and 1.87×10^{44} respectively.

References

Engels, W.R. (2009). Exact Tests for HardyWeinberg Proportions. *Genetics*, **183**,1431 – 1441.

Gail, M. and Mantel, N. (1977). Counting the number of $r \times c$ contingency tables with fixed margins. *Journal of the American Statistical Association*, **72**, 859 – 862.

Zipunnikov, V., Booth, J.G., and Yoshida, R. (2009). Counting tables using the double-saddlepoint approximation. *Journal of Computational and Graphical Statistics*, **18**, 915 – 929.

Constructing a Bivariate Com-Poisson Model Using Copulas

Wasseem Rumjaun[1], Naushad Mamodekhan[2]

[1] University of Mauritius, Reduit, Mauritius
[2] University of Mauritius, Reduit, Mauritius

E-mail for correspondence: wasseem.rumjaun@umail.uom.ac.mu

Abstract: During the past decade, the modeling of multivariate count data has attracted researchers particularly in the field of epidemiology, finance, agriculture and economics. The Conway-Maxwell Poisson distribution, a two-parameter generalisation of the Poisson distribution, is a flexible tool for researchers and statisticians to model under-, equi- as well as over-dispersed counts. Until now, the CMP distribution has been intensively used in univariate cross-sectional and longitudinal regression modeling and till date, no extension of the CMP model in the multivariate set-up has been made. In this paper, we attempt to formulate a Bivariate CMP (BCMP) distribution using the copula techniques. Archimedian copula constructors such as the Clayton, Frank, Gumbel and Ali Mikhail-Haq copulas are used to capture the dependence structure in between the two CMP-Poisson marginals. The parameters of the proposed set ups will be estimated using the Inference for Margins (IFM) principle wherein the parameters of the CMP marginals are estimated separately and the converged estimates are then used to obtain the copula (dependence) parameter via a Newton-Raphson iteration. A simulation study is also devised to test the performance of the BCMP-Copula models. The simulation study concludes that the BCMP-Gumbel model provides the best maximum likelihood estimates for the dependence parameter.

Keywords: COM-Poisson; Copulas; Simulation; Inference for Margin.

1 Conway-Maxwell Poisson

The CMP distribution belongs to the family of exponential densities and is suitable for modelling all types of dispersed data. In several simulation studies and real-life experiments, the CMP distribution has also shown to provide equally good fits as the Negative Binomial or Generalised-Poisson

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

models in the context of over-dispersion. Sellers and Shmueli (2010) developed a Maximum Likelihood Estimation (MLE) for a single link GLM involving the CMP distribution. In terms of the inference procedures, the MLE via the Fisher-Scoring (FS) algorithm adaptation has been used to estimate the mean and dispersion parameters.

1.1 Developing the BCMP-Copula

For a given random variable Y, the CMP's pdf is expressed as:

$$f(y) = \frac{\lambda^y}{(y!)^\nu Z(\lambda,\nu)},\ y = [0,\infty], \lambda > 0, \nu \geq 0 \tag{1}$$

where ν is the dispersion parameter and $Z(\lambda,\nu) = \sum_{s=0}^{\infty} \frac{\lambda^s}{(s!)^\nu}$.
To facilitate the computation of parameters, Minka et al. (2003) suggested the asymptotic approximation of the normalising constant.

$$Z(\lambda,\nu) = \frac{e^{\nu\lambda^{1/\nu}}}{\lambda^{\frac{\nu-1}{2\nu}}(2\pi)^{\frac{\nu-1}{2}}\sqrt{\nu}}[1 + O(\lambda^{-1/\nu})] \tag{2}$$

From which, $E(Y) \approx \lambda^{1/\nu} - \frac{\nu-1}{2\nu}$, $\text{Var}(Y) \approx \frac{\lambda^{1/\nu}}{\nu}$.
We develop a BCMP model based on the closed-form weight structure by using four commonly used copulas in the discrete context: Frank (F), Gumbel (G), Ali-Mikhail Haq (AMH) and Clayton (C).

- **BCMP-C**

$$f^C(y_{1j}, y_{2j}; \kappa) = C_\kappa^C(F_1(y_{1j}), F_2(y_{2j}))f_1(y_{1j})f_2(y_{2j}); \tag{3}$$

$$C_\kappa^C(u_1, u_2; \kappa) = (u_1^{-\kappa} + u_2^{-\kappa} - 1)^{-1/\kappa}, \kappa \in [-1,\infty]$$

- **BCMP-F**

$$f^f(y_{1j}, y_{2j}; \kappa) = C_\kappa^f(F_1(y_{1j}), F_2(y_{2j}))f_1(y_{1j})f_2(y_{2j}); \tag{4}$$

$$C_\kappa^F(u_1, u_2; \kappa) = -\frac{1}{\kappa}\log\left\{1 + \frac{(e^{-\kappa u_1} - 1)(e^{-\kappa u_2} - 1)}{e^{-\kappa} - 1}\right\}, \kappa \in [-\infty,\infty]$$

- **BCMP-G**

$$f^G(y_{1j}, y_{2j}; \kappa) = C_\kappa^G(F_1(y_{1j}), F_2(y_{2j}))f_1(y_{1j})f_2(y_{2j}); \tag{5}$$

$$C_\kappa^G(u_1, u_2; \kappa) = \exp\{-[(-\ln u_1)^\theta + (-\ln u_2)^\kappa]^{1/\kappa}\}, \kappa \in [1,\infty]$$

- **BCMP-AMH**

$$f^{AMH}(y_{1j}, y_{2j}; \kappa) = C_\kappa^{AMH}(F_1(y_{1j}), F_2(y_{2j}))f_1(y_{1j})f_2(y_{2j}); \tag{6}$$

$$C_\kappa^{AMH}(u_1, u_2; \kappa) = \frac{u_1 u_2}{(1 + \kappa(1-u_1)(1-u_2))}, \kappa \in [-1,1]$$

The estimation procedure is carried out via IFM, Joe and Xu (1996), where the BCMP parameters are first obtained under the independence assumption. In the second step, the copula part of the likelihood equations is maximised to obtain the corresponding dependence parameter.

2 Simulation and Results

10,000 pairs of counts are drawn from a Bivariate Normal distribution with $\mu = (0,0)^{\mathrm{T}}$ and a compound symmetrical covariance structure with parameters $\rho = 0.3, 0.5, 0.9$. The pairs are transformed into CMP counts using the Cornish-Fisher expansion. For equi-dispersion, $\lambda_1 = \lambda_2 = 5$ and $\nu_1 = \nu_2 = 1$. For under-dispersion, $\lambda_1 = \lambda_2 = 15$ and $\nu_1 = \nu_2 = 1.1$ while for over-dispersion, $\lambda_1 = \lambda_2 = 4$ and $\nu_1 = \nu_2 = 0.5$. The parameters for over- and under-dispersion have been chosen such that both satisfy the condition $\lambda > 10^{1/\nu}$ for which the approximation (2) becomes accurate. The results including the AIC under each set up are summarised in Table 1.

TABLE 1. MLE of Simulated Data.

ρ	$\hat{\lambda}_1$ $\hat{\lambda}_2$	$\hat{\nu_1}$ $\hat{\nu_2}$	$\hat{\kappa_C}$ $\hat{\kappa_F}$	$\hat{\kappa_G}$ $\hat{\kappa_A}$
0.3	5.589 1.064	5.393 1.045	0.443 1.946	1.180 0.719
			87181 86975	87143 87020
0.5	5.219 1.026	5.208 1.025	0.820 3.423	1.385 NA
			86250 85617	85802 NA
0.9	5.198 1.024	5.132 1.018	3.244 11.155	2.862 NA
			76714 73973	73556 NA
0.3	15.013 1.099	14.896 1.095	0.270 1.731	1.185 0.634
			103729 103393	103468 103450
0.5	14.917 1.098	15.081 1.099	0.628 3.391	1.447 NA
			102293 101365	101344 NA
0.9	15.197 1.104	14.597 1.087	2.863 11.356	3.049 NA
			92791 89528	88664 NA
0.3	4.106 0.504	3.998 0.495	0.270 1.945	1.242 0.675
			126071 125721.1	125720.5 125794
0.5	4.012 0.497	3.911 0.488	0.512 3.273	1.460 NA
			125098 124302	124160 NA
0.9	4.027 0.498	4.027 0.498	2.537 11.664	3.199 NA
			115773 111684	110657 NA

The main remarks from Table 1 are that the simulated values of $\hat{\lambda}_1$, $\hat{\lambda}_2$, $\hat{\nu_1}$ and $\hat{\nu_2}$ are reliable for the chosen sample size of 10,000. The maximum likelihood estimates of the CMP parameters are in fact significant and close to their expected values under each dispersion case.

In the second step, the estimated values of the CMP marginals were used to obtain the Copula (dependence) parameters.
The copula parameters estimated are all significant and within their permissible range. It is noted that values of $\hat{\kappa_A}$ are not obtained for $\rho = 0.5, 0.9$ as the copula parameter will always be out of range when the product-moment correlation is not within (-0.271, 0.478) when marginals are uniform or approximately (-0.300,0.600) for normal marginals - Johnson (1987). From the AIC values reported, the BCMP-G model offered the best estimate across all levels of dispersion for $\rho = 0.5, 0.9$ while the BCMP-F model is most suitable when $\rho = 0.3$.
It is nevertheless observed that the difference in AIC values reported under the Frank and Gumbel copula models are very close to each other.

3 Conclusion

This paper introduces the setting up of Bivariate Conway-Maxwell Poisson models using Copulas. Four BCMP-Copula models namely the BCMP-Clayton, BCMP-Frank, BCMP-Gumbel and BCMP-AMH were proposed and tested via a simulation study designed to generate CMP counts at different levels of dispersion. The BCMP-F model is recommended in the case of low correlation while the BCMP-G set up is most suitable for higher correlations. Aside from the BCMP-AMH under high correlation set-ups, the proposed models, performed well during the simulation study. Through BCMP-Copula models, a more flexible approach towards time series analysis and Bivariate analysis of counts can be undertaken for the future.

Acknowledgments: Special Thanks to Dr. Naushad Mamodekhan for his insights to the completion of the work. I would like to express my gratitude to my parents and friends for their support and indirect contributions to the paper. Much regards to the IWSM for the acceptance of this work.

References

Joe, H. and Xu, J.J. (1996). The estimation method of inference functions for margins for multivariate models. *Technical Report*, **166**, Department of Statistics, University of British Columbia.

Johnson, M (1987). *Multivariate Statistical Simulation.* New York: Wiley.

Minka, T.P, Shmueli, G, Kadane, J.B, Borle, S, Boatwright, P. (2003). Computing with the COM-Poisson distribution. *Technical Report Series*, Carnegie Mellon University Department of Statistics, Pennsylvania.

Sellers, K. and Shmueli, G. (2010). A flexible regression model for count data. *Ann. Appl. Stat*, **4(2)**, 943 – 961.

Tuning parameter selection in LASSO regression

Gianluca Sottile[1], Vito M. R. Muggeo[1]

[1] Università degli Studi di Palermo, Italy

E-mail for correspondence: gianluca.sottile@unipa.it

Abstract: We propose a new method to select the tuning parameter in lasso regression. Unlike the previous proposals, the method is iterative and thus it is particularly efficient when multiple tuning parameters have to be selected. The method also applies to more general regression frameworks, such as generalized linear models with non-normal responses. Simulation studies show our proposal performs well, and most of times, better when compared with the traditional Bayesian Information Criterion and Cross validation.

Keywords: tuning parameter selection; lasso; GCV; BIC; CV; Schall algorithm.

1 Introduction

In the context of high-dimensional data, typically only a small number of variables are truly informative whilst other are redundant. Selecting the appropriate variables is a crucial step of data analysis process. An underfitted model excluding truly informative variables may lead to severe estimation bias in model fit, whereas an overfitted model including redundant uninformative variables, increases the estimated variance and hinders model interpretation.

Among different variable selection methods discussed in literature, penalized regression models have gained popularity and attractiveness. Selection of variables is controlled by the tuning parameter which encourages model sparsity. Well known procedures include the Least Absolute Shrinkage and Selection Operator (LASSO, Tibshirani, 1996), the Smoothy Clipped Absolute Deviation (SCAD, Fan and Li, 2001), the Adaptive LASSO (ALASSO, Zou, 2006), and the Elastic Net (Zou and Hastie, 2005). In penalized regression, the tuning parameters balance the trade-off between model fit and model sparsity, and selecting an appropriate value is the key point. In

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

literature, traditional criteria to select the tuning parameter include Cross-Validation (CV), Generalized Cross-Validation (GCV), Mallows Cp, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).
Broadly speaking, Wang, Li and Tsai (2007b) found that the resulting model selected by GCV tends to overfit, while the BIC is able to identify consistently the finite-dimensional true model. Wang, Li and Tsai (2007b) also indicated that GCV is similar to AIC. Nowadays, many authors proposed to select the tuning parameters through the k-fold CV, which is also the default option in several R packages. The common feature of the aforementioned traditional methods is the grid-search optimization, wherein several candidate tuning parameter values are fixed and different models corresponding to such selected values are fitted.
This article proposes a tuning parameter selection method based on an iterative algorithm which works not only in the classical framework $n > p$ but also in the high-dimensional $n \leq p$ or very high-dimensional setting $n \ll p$. The rest of the article is organized as follows. Section 2 briefly presents the iterative algorithm, Section 3 shows some simulation studies and Section 4 gives some conclusions about the new method proposed.

2 Methods

In LASSO regression with sample size n and p covariates, the objective is to find a solution of the following optimization problem:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{1}$$

As discussed in the Introduction, typically one fixes the tuning parameter λ that balances the trade-off between sparsity and fitting, and then minimizes objective (1) by means of any of optimization algorithms recently developed, e.g. gradient descent or lars.
To set up an iterative algorithm to find λ, we borrow the Schall algorithm idea successfully employed to estimate the variance components in random effects models. More specifically, starting from an initial guess, $\lambda^{(0)} = .001$, say, the algorithm alternates estimation of lasso regression (at fixed λ) and updating of the tuning parameter via the variance ratio properly modified to account for the L_1 penalty.

3 Simulation Studies

Some simulations have been undertaken to compare the traditional selection criteria, BIC, GCV, CV with respect to the proposed algorithm. To assess the performance of each selection criterion, we report degrees of freedom (*df*, the number of non null coefficients), and the mean squared error

(MSE) corresponding to OLS fits including only the informative covariates selected.

TABLE 1. Performance of Tuning Parameter Selector criteria in LASSO Regression (BIC/EBIC, GCV, CV and the proposed algorithm, New): MSE for 2 different sample sizes and several p/n ratios. Results based on 500 simulation runs.

	$n = 50$				$n = 200$			
p/n	(E)BIC	GCV	CV	New	(E)BIC	GCV	CV	New
				$n > p$				
0.15	0.150	0.163	0.171	0.125	0.059	0.107	0.103	0.029
0.25	0.183	0.218	0.218	0.136	0.061	0.142	0.128	0.030
0.35	0.214	0.288	0.283	0.132	0.069	0.170	0.157	0.031
0.45	0.219	0.324	0.308	0.146	0.068	0.199	0.172	0.030
0.55	0.253	0.370	0.361	0.148	0.077	0.225	0.202	0.031
0.65	0.245	0.399	0.375	0.151	0.070	0.233	0.204	0.032
0.75	0.266	0.449	0.403	0.150	0.075	0.261	0.214	0.030
0.85	0.258	0.502	0.412	0.182	0.073	0.269	0.235	0.031
0.95	0.277	0.577	0.422	0.185	0.076	0.284	0.236	0.030
				$n \leq p$				
1	0.303	0.721	0.443	0.204	0.079	0.304	0.241	0.029
2	0.335	0.590	0.566	0.138	0.087	0.347	0.299	0.030
4	0.420	0.757	0.681	0.225	0.088	0.436	0.357	0.030
8	0.477	0.824	0.724	0.264	0.096	0.548	0.031	0.030
16	0.496	0.851	0.778	0.318	0.106	0.632	0.495	0.031
32	0.607	0.922	0.757	0.392	0.121	0.706	0.558	0.031
64	0.844	0.991	0.788	0.434	0.125	0.745	0.564	0.032

TABLE 2. Performance of Tuning Parameter Selector criteria in LASSO Regression (BIC/EBIC, GCV, CV and the proposed algorithm, New). Average degrees of freedom and average number of correctly selected variables (in brackets) for 2 different sample sizes and several p/n ratios. Results based on 500 simulation runs.

	$n = 50$				$n = 200$			
p/n	(E)BIC	GCV	CV	New	(E)BIC	GCV	CV	New
				$n > p$				
0.15	5.7(5)	6.2(5)	6.4(5)	5.1(5)	6.3(5)	11.1(5)	10.8(5)	5.0(5)
0.25	6.2(5)	7.3(5)	7.6(5)	5.2(5)	6.1(5)	12.8(5)	11.6(5)	5.0(5)
0.35	6.5(5)	8.9(5)	8.8(5)	5.1(5)	6.4(5)	14.1(5)	13.1(5)	5.0(5)
0.45	6.6(5)	9.6(5)	9.2(5)	5.4(5)	6.2(5)	15.8(5)	13.5(5)	5.0(5)
0.55	7.2(5)	10.9(5)	10.5(5)	5.5(5)	6.4(5)	17.4(5)	15.0(5)	5.0(5)
0.65	6.6(5)	11.2(5)	10.4(5)	5.3(5)	6.0(5)	17.2(5)	14.4(5)	5.0(5)
0.75	7.0(5)	12.7(5)	11.0(5)	5.3(5)	6.3(5)	19.4(5)	15.1(5)	5.0(5)
0.85	6.7(5)	14.7(5)	10.9(5)	5.7(5)	6.1(5)	19.5(5)	16.0(5)	5.0(5)
0.95	7.1(5)	18.8(5)	11.1(5)	5.9(5)	6.2(5)	20.5(5)	15.5(5)	5.0(5)
				$n \leq p$				
1	7.1(5)	26.4(5)	11.1(5)	5.8(5)	6.4(5)	22.6(5)	16.0(5)	5.0(5)
2	7.5(5)	14.2(5)	13.7(5)	5.1(5)	6.3(5)	21.1(5)	17.6(5)	5.0(5)
4	8.6(5)	19.6(5)	17.0(5)	5.9(5)	6.2(5)	25.7(5)	20.0(5)	5.0(5)
8	9.2(5)	22.1(5)	18.2(5)	6.4(5)	6.3(5)	34.4(5)	23.9(5)	5.0(5)
16	9.0(5)	23.3(5)	19.3(5)	6.6(5)	6.4(5)	41.9(5)	27.9(5)	5.0(5)
32	11.0(5)	28.0(5)	18.1(5)	7.3(5)	6.6(5)	49.0(5)	32.8(5)	5.0(5)
64	20.2(5)	37.3(5)	19.2(5)	9.6(5)	6.5(5)	54.3(5)	31.1(5)	5.0(5)

The simulated data come from $y = X\beta + \epsilon$, where $X \sim N_p(0_p, \Sigma_p)$, ($\Sigma_{jk} = 0.5^{|j-k|}$) and $\epsilon \sim N(0,1)$. For two sample sizes (50, 200), two different scenarios have been considered: in the first scenario ($n > p$), $p \in \{.15n, .25n, \ldots, .95n\}$ and true coefficients $\beta = (5, 4, 3, 2, 1, 0, \ldots, 0)^{\mathrm{T}}$. In the second scenario, $n < p$, $p \in \{1n, 2n, \ldots, 64n\}$ and β as in the first scenario.

Table 1 and 2 report average mean squared errors (MSE), average degrees of freedom (*df*) and the number of correctly selected parameters. Results show that the proposed method performs better than the others in all the scenarios, not only in terms of model fit but also in terms of degrees of freedom.

The proposed iterative algorithm always exhibits the lowest MSE, but when $n < p$, particularly with small samples ($n = 50$), the new methods performs largely better than the other competitors.

4 Conclusions

We have introduced a 'new' approach to select iteratively the tuning parameter λ of lasso regression. Limited simulation evidence suggests the method attains comparatively better performance in all considered settings. Results have been presented for the classical Gaussian model, but the proposed approach is favored to be employed in generalized linear models with binary or count responses. Application in very high-dimensional settings ($n \ll p$) that are today one of the most challenging concerns, represents a noteworthy point to be investigated.

References

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58** 267 − 288.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, **96**, 1348 − 1360.

Wang, H., Li, R., and Tsai, C.L. (2007b). Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method. *Biometrika*, **94**, 553 - 568.

Longitudinal models with informative time measurements

Inês Sousa[1], Ana Borges[2], Luis Castro[3]

[1] Centro de Matemática, University of Minho, Portugal
[2] Escola Superior de Tecnologia e Gestão de Felgueiras, Instituto Politécnico do Porto, Portugal
[3] Hospital de Braga, Portugal

E-mail for correspondence: isousa@math.uminho.pt

Abstract: In longitudinal studies individuals are measured repeatedly over a period of time for a response variable of interest. In classical longitudinal models the longitudinal observed process is considered independent of the times when measurements are taken. However, in medical context it is common that patients in worst health condition are more often observed, whereas patients under control do not need to be seen so many times. Therefore, longitudinal models for data with this characteristics should allow for an association between longitudinal and time measurements processes. In this work we propose a joint model for the distribution of longitudinal response and time measurement using likelihood inference for the model parameters. A simulation study is conducted and the model proposed is fitted to a data set on progression of oncological biomarkers in breast cancer patients.

Keywords: longitudinal; follow-up times; biomarkers

1 Introduction

In longitudinal studies individuals are measured repeatedly over a period of time. In unbalanced observational studies individuals have different number of measurements assessed at different times. Usually, in medical context, patients are observed for a response variable of interest up to doctors decision rather then based on a predefined protocol. That is, patients are usually measured according to their clinical condition. For example, when monitoring for breast cancer progression, patients are repeatedly measured for biomarkers CEA and CA15.3 based on their historical measurements.

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

Therefore, in these cases the follow-up time process should be considered dependent on the longitudinal outcome process.
The general linear model (Diggle et al. 2002) described for longitudinal data analysis, assumes a deterministic follow-up time process that is noninformative about the outcome longitudinal response. In this work we propose to joint model the longitudinal process and the follow-up time process conditional on the historical unobserved longitudinal process.
Others have been proposed models for situations where the longitudinal response variable and the time measurements are related. More lately, Fang et al (2016) proposed a joint model for longitudinal and informative observation using two random effects with additive mixed effect model for observation time. Cheng et al (2015) proposed a model where the probability structure of the observation time process is unspecified. Lipsitz et al (2002) consider a model where assumptions regarding the time measurements process result in the likelihood function separated in the two components. Lin et al (2004) approach is base on missing data and proposed a class of inverse intensity-of-visit process-weighted estimators in marginal regression models. Fitzmaurice et al (2006) consider the same problem when the longitudinal response is binary.
In this work we consider a response longitudinal variable with Gaussian distribution. We propose a model where the follow-up time process is stochastic. The model is described through the joint distribution of the observed process and the follow-up time process. Estimation of model parameters is through maximum likelihood, where a Monte Carlo approximation is necessary. We conducted a simulation study of longitudinal data where model parameter estimates are compared, when using the model proposed and ignoring the association between processes. Finally, the model proposed is applied to a real data set on progression of oncological biomarkers in breast cancer patients.

2 Model Proposal

Consider data observed for m individuals, where $\mathbf{Y}_i$ is the vector of longitudinal responses and $\mathbf{T}_i$ is the vector of time measurements, both for subject $i = 1, ..., m$. It is assumed a model for the joint distribution of the longitudinal outcome process $\mathbf{Y}$ and the time measurement process $\mathbf{T}$ through an unobserved stationary Guassian process $\mathbf{W}(\cdot)$. Therefore, we propose the following model

$$[\mathbf{Y}_i|\mathbf{W}(\mathbf{T}_i), \mathbf{T}_i] \sim \text{Normal}(\mu + \mathbf{W}(t_{ij}), \tau^2)$$

and intensity function for the time measurement process at time t_{ij}, $j = 1, ..., n_i$

$$\lambda(\mathbf{T}_{ij})|\mathbf{W}_{history}(s) \sim \exp\left\{\mathcal{F}(\mathbf{W}_{history}(t_{ij}))\right\},$$

where, μ is the expected value that can include regression parameters and $\mathcal{F}(.)$ is any defined function. For example, to describe a time measurement process dependent on the progression of the patients unobserved health condition, we might define

$$\lambda(\mathbf{T}_{ij})|\mathbf{W}_{history}(s) = \exp\left(\alpha + \gamma \int_0^{t_{ij}} W(s)ds\right).$$

Notice that, process $\mathbf{W}(\cdot)$ is continuous in time, though only a discrete version of it is observed at t_{ij}.,

For inference we consider a likelihood approach, where the likelihood function is

$$\begin{aligned}
[\mathbf{Y},\mathbf{T}] &= \prod_{i=1}^{m}[\mathbf{Y}_i,\mathbf{T}_i] \\
&= \prod_{i=1}^{m}\int_W [\mathbf{Y}_i|\mathbf{W}][\mathbf{T}_i|\mathbf{W}][\mathbf{W}]dW \\
&= \prod_{i=1}^{m} E_{\mathbf{W}|\mathbf{Y}_i}\left([\mathbf{T}_i|\mathbf{W}][\mathbf{Y}_i|\mathbf{W}_0]\frac{[\mathbf{W}_0]}{[\mathbf{W}_0|\mathbf{Y}_i]}\right)
\end{aligned}$$

where, $\mathbf{W}_0$ is the subset with observed time points and $\mathbf{W}_1$ is the subset withunobserved time points.,

We then generate g samples from $[\mathbf{W}|\mathbf{Y}_i$ and approximate the expectation by its Monte Carlo version

$$L_{MC}(\theta) = \prod_{i=1}^{m}\frac{1}{g}\sum_{j=1}^{g}\left(f(\mathbf{T}_i|\mathbf{W}_j)f(\mathbf{Y}_i|\mathbf{W}_{0j})\frac{f(\mathbf{W}_{0j})}{f(\mathbf{W}_{0j}|\mathbf{Y}_i)}\right)$$

3 Results

A simulation study is conducted and results are presented when fitting the model proposed and the general linear longitudinal model (Diggle et al, 2002).,

A data set on oncological biomarkers, CEA and CA15.3, for breast cancer patients is available from Hospital de Braga, Portugal. There are data available on 550 patients, with a mean number of measurements per subject of 7.6 (median=7 and sd=4.1), with a total number of observations for CEA of 4166 and 5166 for CA15.3. In Figure 1 longitudinal profiles of CEA and CA15.3 (logarithm scale) of a random sample of 10 patients is shown, with black dots representing the location of the time measurements and the solid black line is the respective smooth spline for all data.

The proposed model is fitted to this data and results are compared with the classical longitudinal model.

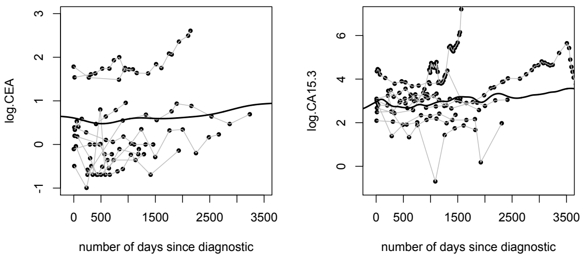


FIGURE 1. Longitudinal profiles of a random sample of 10 patients measured for CEA and CA15.3.

References

Chen, Y., Ning, J. and Cai, C.Y. (2015). Regression analysis of longitudinal data with irregular and informative observation times. *Biostatistics*, **16**, 727 – 739.

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

Fang, S., Zhang, H.X. and Sun, L.Q. (2016). Joint analysis of longitudinal data with additive mixed effect model for informative observation times. *Journal of Statistical Planning and Inference*, **169**, 43 – 55.

Fitzmaurice, G., Lipsitz, S., Ibrahim, J., Gelber, R. and Lipshultz, S. (2006). Estimation in regression models for longitudinal binary data with outcome-dependent follow-up. *Biostatistics*, **7**, 469 – 485.

Lin, H., Scharfstein, D. and Rosenheck, R. (2004). Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of the Royal Statistical Society, Series B*, **66**, 791 – 813.

Lipsitz, S., Fitzmaurice, G., Ibrahim, J., Gelber, R. and Lipshultz, S. (2002). Parameter estimation in longitudinal studies with outcome-dependent follow-up. *Biometrics*, **58**, 621 – 630.

Ensemble Post-Processing over Complex Terrain Using High-Resolution Anomalies

Reto Stauffer[12], Nikolaus Umlauf[1], Jakob W. Messner[12], Georg J. Mayr[2], Achim Zeileis[1]

[1] Department of Statistics, Univ. of Innsbruck, Austria
[2] Institute of Atmospheric and Cryospheric Sciences, Univ. of Innsbruck, Austria

E-mail for correspondence: reto.stauffer@uibk.ac.at

Abstract: Probabilistic weather forecasts computed by numerically solving physical equations describing atmospheric processes have systematic errors, particularly over complex terrain. Statistical post-processing is often applied to alleviate these errors. We will present a novel fully scalable full-distributional post-processing method for precipitation, using high-resolution local anomalies to account for the high spatial variability. The application of the new method to the central Alps improves the skill of forecasts for both the probability of occurrence and the amount of precipitation.

Keywords: GAM; BAMLSS; Censored Normal Distribution; Precipitation; Climatology and Weather.

1 Introduction & Data

In mountainous regions, large amounts of precipitation can lead to severe floods and land slides during spring and summer, and to dangerous avalanche conditions during winter. An accurate and reliable knowledge about the expected precipitation can therefore be crucial for strategical planning and to raise awareness among the public.
Precipitation forecasts are typically provided by numerical weather prediction (NWP) models using physical prognostic equations. Ensemble prediction systems (EPS) provide several independent weather forecasts based on slightly different initial conditions to depict the forecast uncertainty. A crucial limitation of these forecasts is the horizontal resolution. Therefore several approaches to correct the NWP forecasts for unresolved features and systematic errors are available, known as post-processing methods.

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

We present a novel spatial post-processing method for precipitation over complex terrain using high-resolution spatial climatologies as background information, and apply it to Tyrol, Austria. Due to the local alpine topography, observations vary strongly across the domain, increasing the complexity for spatial modelling. The new approach uses high-resolution climatologies to remove local features from (i) the observations and also from (ii) EPS forecasts. The remaining short-term derivations can be used to create high-resolution spatial corrected EPS forecasts.

We use an ensemble consisting of 50 forecasts computed by the EPS of the European Center for Medium-Range Weather Forecasts (ECMWF). The horizontal mesh of the current model is roughly 32 *km* (see Figure 1, left). Approximately 2200 single days (2010–2015) are used, including 90 grid points from the EPS model covering the area of interest. Precipitation observations at 117 stations cover the period 1971–2013 and constitute roughly 1.5 million unique observations.

2 Censored Spatio-Temporal Anomaly Model

Ensemble model output statistics (EMOS; Gneiting et al., 2005) model the statistical relationship between past observations and the corresponding EPS forecasts. As the EPS provides 50 individual forecasts, the corrections can be accounted to both, the expected mean, and the uncertainty of the EPS, typically represented by the EPS standard deviation.

$$y \sim \mathcal{N}\big(\mu, \sigma\big) \quad \text{with:} \;\; \mu = \beta_0 + \beta_1 m(\text{eps}), \;\; \sigma = \gamma_0 + \gamma_1 s(\text{eps}) \tag{1}$$

Gneiting et al. (2005) proposed that the response y is assumed to follow a normal distribution with location μ represented by a linear function of the EPS mean (m(eps)), and standard deviation σ represented by a linear function of the EPS standard deviation (s(eps)).

However, for the application of high-resolution precipitation post-processing on a daily time scale, two major problems arise. Daily sums of observed precipitation are *no longer normally distributed*, as they contain a large fraction of zero-observations (dry days), and the observations show a *large variability across the area of interest* – especially over complex terrain like e.g., the Alps.

To account for the distribution of the observations, the conditional response distribution in Equation 1 has to be modified first. Messner et al. (2014) showed that the response distribution of precipitation can be seen as left-censored normal, as precipitation is physically limited to 0 *mm*.

Furthermore, a way has to be found to include the information of all available stations within the area of interest, but to account for the different location and season dependent characteristics across the domain at the same time. Therefore we are using the concept of local standardised anomalies, based on high-resolution precipitation climatologies. Both, the observations and all 50 individual EPS forecast members, will therefore be standardised using:

$$y^* = \frac{y - \mu_{y,clim}}{\sigma_{y,clim}} \tag{2}$$

While the climatological location $\mu_{y,clim}$ and scale $\sigma_{y,clim}$ represent the long-term spatio-temporal patterns in both, the observations and the individual EPS forecast members respectively (y), anomalies are the short-term deviations from the underlying climatology. By removing location and season dependent characteristics, the observations and the EPS forecasts can be brought to a compareable level, what will be called "standardised anomalies", denoted by superscript "*".

We are using a Bayesian framework estimating generalized additive models for the climatologies (R package `bamlss`, Umlauf et al., 2016) to estimate heteroscedastic spatio-temporal climatologies of the observations, and the EPS forecasts. Therefore, similar assumptions to Equation 1 will are used, replacing the linear predictors for μ, and $\log(\sigma)$ by (Stauffer et al., 2015):

$$\beta_0 + \beta_1 \text{alt} + s(\text{yday}) + s(\text{lon}, \text{lat}) + s(\text{yday}, \text{lon}, \text{lat}) \tag{3}$$

The linear predictor includes a linear altitudinal, a cyclic seasonal ($s(\text{yday})$), a 2-D spatial ($s(\text{lon}, \text{lat})$), and a 3-D effect ($s(\text{yday}, \text{lon}, \text{lat})$) to account for changes in the seasonal pattern across the area of interest. Once the climatologies are known, the statistical relationship between standardised anomalies of the observations, and the standardised anomalies of the EPS forecasts can be modelled similar to the EMOS approach in Equation 1 using:

$$y^* \sim \mathcal{N}\big(\mu, \sigma\big) \text{ with: } \mu = \beta_0 + \beta_1 m(\text{eps}^*),\ \log(\sigma) = \gamma_0 + \gamma_1 \log(s(\text{eps}^*)) \tag{4}$$

As the standardised anomalies are no more location dependent, the prediction for any location within the area of interest can be made. This allows for a spatial correction of any future EPS forecast on an arbitrary fine horizontal resolution.

3 Summary

The novel approach for precipitation using anomalies provides an attractive and reliable new method for spatial ensemble post-processing. Once the climatologies are estimated, the computational costs are very low. Regarding the full probabilistic response, several quantities can be derived from one single model, like the expected amount of precipitation, quantiles, or probabilities. Figure 1 shows spatial sample prediction on a 800 m grid for a +30 h forecast, comparing the raw EPS mean (left) against the corrected forecasts (middle). In contrast to the EPS, several topographical features can be identified after the correction. Beside, probabilities for exceeding two different thresholds are plotted. First results have shown that the novel approach applied to the area of Tyrol, located in the Eastern Alps, increases the forecast skill for both, the probabilities of exceeding a certain threshold, and the amount of precipitation.

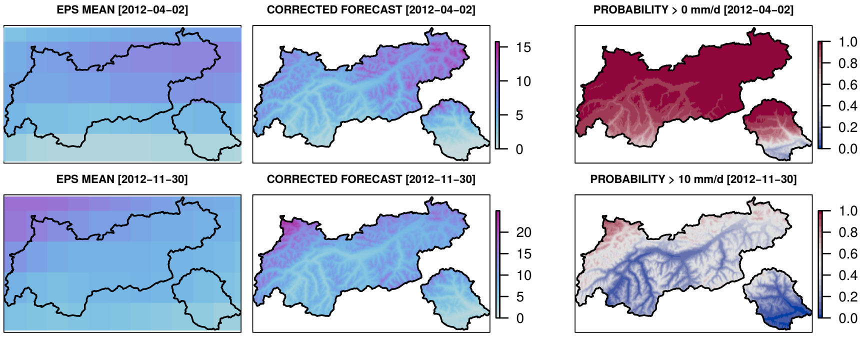

EPS MEAN [2012−04−02]
CORRECTED FORECAST [2012−04−02]
15
10
5
0
PROBABILITY > 0 mm/d [2012−04−02]
1.0
0.8
0.6
0.4
0.2
0.0
EPS MEAN [2012−11−30]
CORRECTED FORECAST [2012−11−30]
20
15
10
5
0
PROBABILITY > 10 mm/d [2012−11−30]
1.0
0.8
0.6
0.4
0.2
0.0

FIGURE 1. Sample predictions. Top: 2012-04-02, bottom: 2012-11-30. Left to right: raw uncorrected EPS forecast $[mm\ d^{-1}]$, corrected forecast $[mm\ d^{-1}]$, and probability of occurrence. Top $> 0\ mm\ d^{-1}$, bottom $> 10\ mm\ d^{-1}$. The color scale for the uncorrected and corrected forecast is identical for each individual day.

Acknowledgments: Ongoing project funded by the Austrian Science Fund (FWF): TRP 290-N26. The computational results presented have been achieved in part using the Vienna Scientific Cluster (VSC). Observation data set provided by the hydrographical service Tyrol (`ehyd.gv.at`).

References

Gneiting T., et. al. (2005): Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review*, **133**, 1098–1118.

Messner J. W., et. al. (2014): Extending Extended Logistic Regression: Extended versus Separate versus Ordered versus Censored. *Monthly Weaather Reveview*, **142**, 3003–3014.

Stauffer, R., et. al. (2015). Spatio-temporal Censored Model of Precipitation Climatology. In: *Proceedings of the International Workshop of Statistical Modelling 2015*, Linz, Austria, 366–371.

Umlauf, N., et al. (2016): `bamlss`: Bayesian Additive Models for Location Scale and Shape (and Beyond). R *package version* `0.1-1` (https://R-Forge.R-project.org/R/?group_id=865).

Bayesian Meta-Analysis to incorporate Uncertainty

Elizabeth Stojanovski

[1] School of Mathematical and Physical Sciences, University of Newcastle, Australia

E-mail for correspondence: elizabeth.stojanovski@newcastle.edu.au

Abstract: The disparate nature of the results from studies for consideration in a meta-analysis may arise due to various aspects including differences in study design, location, method of analysis, among other factors. These differences can be acknowledged and explored through the addition of hierarchies to the meta-analysis model, as in the Bayesian random effects model presented. The ease in with which additional hierarchical levels between study-specific parameters and the overall distribution can be incorporated in the Bayesian framework is demonstrated

Keywords: Meta-analysis; Bayesian ; Between study variation

1 Introduction

A relationship between life events and physical illness has been recognised [Cooper and Payne, 1991]. Information about the associations between particular types of cancer and particular types of life events, which are in effect particular types of life stresses, is sparse and in many cases conflicting [Chen et al. 1995]. Some of this conflict may be attributed to differe bnces in study design characteristics and the controlling of confounding factors. The purpose of a recent random-effects meta-analysis by Duijts et al. [2003] was to identify studies that examined the association between adverse events and the risk of breast cancer and were published to establish the relationship for various types of life events. A group of life events assessed was those associated with financial status changes, which produced an overall non-statistically significant association with breast cancer risk [Roberts et al. 1996] (SOR=0.90, 95% CI: 0.54-1.50). The studies, however, differed substantially in terms of study design, location, time frame and sample size. It

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

is of interest to pool information from various studies, in order to identify characteristics that differentiate study results.
A random-effects Bayesian meta-analysis model is conducted to combine the reported estimates of the four studies described that assess the relationship between life events related to financial status. The proposed model allows three major sources of variation to be taken into account. These include study level characteristics, between study variance and within study variance. The sensitivity of the overall results to various study characteristics is also investigated.

2 Methods

A summary of study-specific characteristics of the studies considered for the meta-analysis is provided in Tables 1 and 2. The study-specific estimates of the association between breast cancer and life events related to financial status are presented in Table 3.
The study characteristics that were considered under Model 2 were as follows: C1: Study design: case control or cohort; C2: Study location: USA or UK; C3: Correction for confounding: yes or no. Thus for each of these three situations, each true study-specific log odds ratio arises from one of two subgroups with subgroup mean log odds ratio.

TABLE 1. Description of studies considered for the meta-analysis.

Study	Author	Time frame	Year of Publication	Country	Design	Exposition
1	Roberts	5 years	1996	USA	Retrospective case-control	Questionnaire
2	Cooper	2 years	1989	UK	Prospective case-control	Questionnaire
3	Cooper	2 years	1993	UK	Limited Prospective cohort	Questionnaire
4	Snell	5 years	1971	USA	Retrospective case-control	Interview

Independence is assumed between studies with this model, so that the precision matrices are all diagonal. The prior precision matrices are defined to have diagonal entries equal to one, reflecting little information and therefore strong uncertainty about between-study variation.
Bayesian analysis involves integration over potentially high dimensional probability distributions of model parameters to enable inferences about

TABLE 2. Description of studies considered for the meta-analysis (continued).

Study	Number of cases	Source of cases	Age of cases	Number of Controls	Source of Cohort	Correction for confounding
1	258	Population	64.8	614	Population	Yes
2	171	Suspicion	55	1992	Hospital	Yes
3	171	Suspicion	55	727	Suspicion	Yes
4	352	Hospital	55.5	670	Hospital	No

model parameters [Fryback et al. 2001]. MCMC may be used instead to draws samples from the required distributions and then form sample averages to approximate expectations. The analyses was undertaken in WinBUGS [Speigelhalter et al. 2000], with a burn-in of 100,000 iterations which are excluded, and a collection period of 100,000 iterations to estimate the odds of developing breast cancer as a result of having experienced life events related to financial status; and initial values were set at the maximum likelihood values. These were more than sufficient to confirm convergence, as indicated by the diagnostics within WinBUGS.

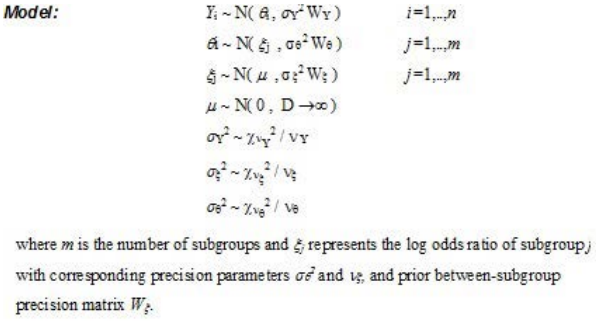
TABLE 3. Study-specific estimates used in the Meta-analysis.

Study	Odds Ratio	95%CI	Log odds ratio	Precision
1	0.96	0.66-1.41	-0.0408	27.77
2	0.65	0.44-0.96	-0.4308	26.29
3	0.59	0.41-0.85	-0.5276	30.10
4	1.73	1.26-2.36	0.5481	40.63

Summaries of the posterior distributions were assessed graphically using kernel density plots and are presented numerically by calculating summary statistics such as the mean, variance and quantiles of the sample. WinBUGS trace and history functions offer serial plots of the actual sequence of simulated values to diagnose convergence. The full empirical distribution function is used for hypothesis testing.

3 Results

The model was employed to inspect the impact of various study design characteristics that differed between the four studies. Trace plots and posterior



Model:

$Y_i \sim N(\theta_i, \sigma_Y^2 W_Y)$ $i=1,..,n$

$\theta_i \sim N(\xi_j, \sigma_\theta^2 W_\theta)$ $j=1,..,m$

$\xi_j \sim N(\mu, \sigma_\xi^2 W_\xi)$ $j=1,..,m$

$\mu \sim N(0, D \rightarrow \infty)$

$\sigma_Y^2 \sim \chi_{\nu_Y}^2 / \nu_Y$

$\sigma_\xi^2 \sim \chi_{\nu_\xi}^2 / \nu_\xi$

$\sigma_\theta^2 \sim \chi_{\nu_\theta}^2 / \nu_\theta$

where m is the number of subgroups and ξ_j represents the log odds ratio of subgroup j with corresponding precision parameters σ_θ^2 and ν_ξ, and prior between-subgroup precision matrix W_ξ.

FIGURE 1. Model

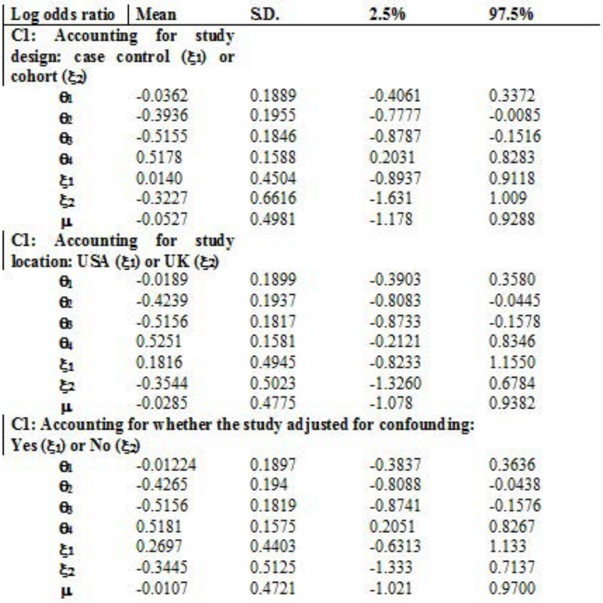
density plots for the model parameters were inspected for stability and conformity to the anticipated distributions. In all cases, these characteristics were confirmed. Estimates of the posterior mean, standard deviation and 95% credible interval for θ, ξ, μ, under each of the three alternatives C1, C2, C3 are given in Table 4. These results suggest that the overall odds ratio from the three case control studies is greater than unity and that from the cohort study is less than unity, although both estimates have 95% credible intervals that span unity. Similarly, those studies conducted in the USA have an overall odds ratio that is greater than unity whereas those conducted in the UK have an overall odds ratio that is less than unity, but again the two credible intervals both include unity. Finally, the overall odds ratio for the three studies that controlled for confounding is greater than unity compared to a reduced odds ratio for the study that did not control for such issues. The overall odds ratios for the three analyses are not substantially different in light of the very wide credible intervals which are a consequence of the disparate study estimates and vague priors.

4 Discussion

By allowing for differences in study design the present analysis supports the findings from the study by Duijts et al. [2003], where it was concluded that life events related to changes in financial status are not statistically significantly related to breast cancer. The disparate nature of the results from these four studies may arise because of differences in study design, location, method of analysis, among other factors. These differences can be acknowledged and explored through the addition of hierarchies to the meta-analysis model, as in the model presented. Because of the small number of studies, the analyses under this model are intended to be indicative rather than substantive.

Unfortunately, there is insufficient information to further investigate these suggested differences in odds ratios associated with different study design

TABLE 4. Summary statistics for the posterior mean log odds ratios θ and μ



Log odds ratio	Mean	S.D.	2.5%	97.5%
C1: Accounting for study design: case control (ξ_1) or cohort (ξ_2)				
θ_1	-0.0362	0.1889	-0.4061	0.3372
θ_2	-0.3936	0.1955	-0.7777	-0.0085
θ_3	-0.5155	0.1846	-0.8787	-0.1516
θ_4	0.5178	0.1588	0.2031	0.8283
ξ_1	0.0140	0.4504	-0.8937	0.9118
ξ_2	-0.3227	0.6616	-1.631	1.009
μ	-0.0527	0.4981	-1.178	0.9288
C1: Accounting for study location: USA (ξ_1) or UK (ξ_2)				
θ_1	-0.0189	0.1899	-0.3903	0.3580
θ_2	-0.4239	0.1937	-0.8083	-0.0445
θ_3	-0.5156	0.1817	-0.8733	-0.1578
θ_4	0.5251	0.1581	-0.2121	0.8346
ξ_1	0.1816	0.4945	-0.8233	1.1550
ξ_2	-0.3544	0.5023	-1.3260	0.6784
μ	-0.0285	0.4775	-1.078	0.9382
C1: Accounting for whether the study adjusted for confounding: Yes (ξ_1) or No (ξ_2)				
θ_1	-0.01224	0.1897	-0.3837	0.3636
θ_2	-0.4265	0.194	-0.8088	-0.0438
θ_3	-0.5156	0.1819	-0.8741	-0.1576
θ_4	0.5181	0.1575	0.2051	0.8267
ξ_1	0.2697	0.4403	-0.6313	1.133
ξ_2	-0.3445	0.5125	-1.333	0.7137
μ	-0.0107	0.4721	-1.021	0.9700

characteristics, or to identify whether there are interactions between these study characteristics. However, the analyses do serve to demonstrate the application of a random-effects Bayesian meta-analysis model by combining results from studies while accommodating partial exchangeability between studies, acknowledging that some studies are more similar due to common designs, locations and so on. The ease in with which additional hierarchical levels between study-specific parameters and the overall distribution can be incorporated in the Bayesian framework is also demonstrated.

References

Chen C.C., David A.S., Nunnerley H., Michell M., Dawson J.L., Berry H., Dobbs J. and Fahy (1995). Adverse life events and breast cancer: case-control study. *British Medical Journal* 311,1527 – 1530.

Cooper C.L. and Payne R. (1991). *Personality and stress: individual differences in the stress process*. Chichester, UK: Wiley.

Duijts S., Zeegers M. and Borne B. (2003). The association between stressful life events and breast cancer risk: A meta-analysis. *International Journal of Cancer* 107,1023 – 1029.

Fryback D., Stout N. and Rosenberg M. (2001). An Elementary Introduction to Bayesian Computing Using WinBUGS. *International Journal of Technology Assessment in Health Care* 17, 98 – 113.

Roberts F., Newcomb P., Trentham-Dietz A., Storer B. (1996). Self-Reported Stress and Risk of Breast Cancer *Cancer*, 77, 1089 – 1093.

Spiegelhalter D.J., Thomas A. and Best N. (2000). Sire evaluation and genetic trends. In: *WinBugs Version 1.4 User Manual.* Software available at http://www.mrcbsu.cam.ac.uk/bugs/winbugs/contents.shtml.

Nonlinear modelling of experimental rock failure data

Daniel Tait[1], Nadine Geiger[1], Bruce J. Worton[1], Andrew Bell[2], Ian G. Main[2]

[1] School of Mathematics and Maxwell Institute for Mathematical Sciences, The University of Edinburgh, Edinburgh, UK
[2] School of GeoSciences, The University of Edinburgh, Edinburgh, UK

E-mail for correspondence: Bruce.Worton@ed.ac.uk

Abstract: In this paper we investigate nonlinear models for rock failure data. The complexities of estimation and analysis using sample experimental data are studied when fitting a power-law type model. Both the high frequency of sampling of observations during the experiment, as well as the nature of the model with high levels of correlation of the parameter estimators result in some challenging issues in the modelling.

Keywords: Nonlinear regression; Least squares estimation; Likelihood; Power-law model.

1 Introduction

We consider methods for modelling experimental data collected in the lab concerning rock failure. Accelerating rates of foreshocks are often observed precursory to natural hazards such as earthquakes and volcanic eruptions. Similarly, rock failure in laboratory experiments is preceded by accelerating strain rates. In this work we investigate the usage of a damage mechanics model proposed by Main (2000) for the analysis of strain and strain rate data during the tertiary phase of brittle creep. The model studied consists of 3 parameters, of which we focus on the failure time and on the power-law exponent. When examining the likelihood function and the Fisher Information we find that there is substantial correlation between these parameters.

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

2 Model

Main (2000) develops a damage mechanics model to explain the time-dependent, trimodal behaviour of brittle creep. We consider modelling the strain using a relationship of the form

$$\Omega = \Omega_I(1 + \frac{t}{m\tau_1})^m + \Omega_{III}(1 - \frac{t}{t_f})^{-v},$$

where for time t, Ω is the strain. The parameters are Ω_I, Ω_{III}, m, τ_1, v and t_f. There is also interest in modelling strain rate $\dot{\Omega}$, the derivative of Ω with respect to time t. Here we focus on the accelerating crack growth which is associated with the second term on the right-hand-side of the above equation as it illustrates the complexities of modelling the data, i.e. a strain model of the form

$$\Omega = \omega(1 - \frac{t}{t_f})^{-v}, \tag{1}$$

where t_f represents the time of failure which is of particular interest, while the exponent parameter v relates to the curvature of the strain relationship.

3 Estimation

We assume for the moment that the strain observations are subject to iid experimental errors with variance σ^2, and the expectation at time t has the form given in (1) which leads to a nonlinear model (Bates and Watts, 1988; Fahrmeir et al., 2013).

We apply a nonlinear least squares estimation procedure to fit the model with parameters v and ω for given t_f to the strain data (Heap et al. (2009) gives details of lab experiments). This seems to be an effective numerical approach for parameter estimation. Models were fitted for failure times t_f over a suitable range of values and the best model selected; this corresponded to $t_f = 138.864$. Table 1 presents the estimates of the fitted model with this particular failure time value.

TABLE 1. Estimation of v and ω from strain for $t_f = 138.864$.

Parameter	Estimate	Std. Error
v	2.147×10^{-2}	4.104×10^{-5}
ω	1.741	1.143×10^{-4}
σ	0.001628	

The fitted model and residuals are given in Figure 1. Looking at the left panel we can see that the fitted model appears to represent the experimental data well, but the residuals show some concerning departures from the assumed model. Firstly there are multiple small waves with a length of about ten minutes, and secondly there are two irregular large waves, which indicate a more severe discrepancy between the fitted values and the data. A partial autocorrelation analysis of the residuals suggests that residuals are autocorrelated, which may possibly be due to the nature of the experiment. Therefore, we expect the estimates to be reliable, but the SEs may be misleading.

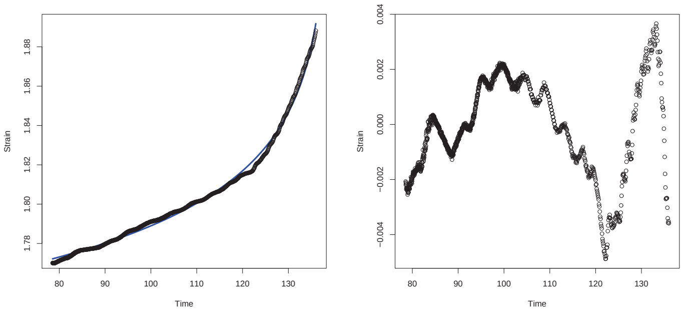


FIGURE 1. Fitted regression line with strain observations (left) and residuals (right) for the estimation from the experimental lab data.

The estimated parameters (obtained by minimising RSS over a range of t_f values, with $\operatorname{argmin}(t_f) = 143.37$) are given in Table 2 for the strain rate data, and the fitted model is shown in Figure 2, with the residuals. The estimates for the strain and the strain rate produce very similar fits when viewed on the strain scale. However, the SEs for the strain rate would appear to be more reliable. On investigating the differences between the parameters estimated by the two models, the t_f over which the RSS was minimised is not very well determined as the RSS does not vary greatly over a range of t_f-models. Also, the parameter estimators are highly correlated so changes in t_f lead to changes in the other two parameters. Nevertheless, the strain and strain rate models provide a useful representation of the experimental data.

TABLE 2. Estimation of v and ω' (the constant multiplicative parameter from $\dot{\Omega}$ model) from strain rate for $t_f = 143.37$.

Parameter	Estimate	Std. Error
v	0.25634	0.05680
ω'	0.13246	0.04688
σ	0.002479	

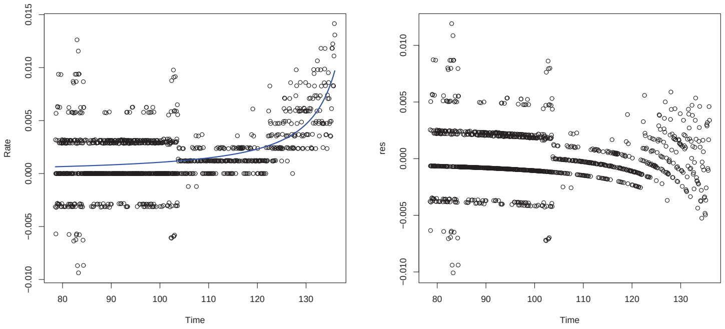


FIGURE 2. Fitted regression line with strain rate observations (left) and residuals (right) for the estimation from the experimental lab data.

4 Discussion

In this paper we have considered the application of nonlinear models for analysis of experimental rock failure data during the tertiary phase of brittle creep. These 3-parameter models have been used to explore different features of the data and highlight some of the challenges of analysing such data. In future work, where lab experiments are repeated under identical conditions, it is expected that the use of replication will enable the features we have seen to be investigated more fully, and provide further insights into the processes related to natural hazards such as earthquakes and volcanic eruptions.

Acknowledgments: Daniel Tait is supported by an EPSRC studentship.

References

Bates, D.M. and Watts, D.G. (1988). *Nonlinear Regression Analysis and its Applications.* New York: Wiley.

Fahrmeir, L., Kneib, T., Lang, S., Marx, B. (2013). *Regression: Models, Methods and Applications.* Heidelberg: Springer.

Heap, M.J., Baud, P., Meredith, P.G., Bell, A.F., and Main, I.G. (2009). Time-dependent brittle creep in Darley Dale sandstone. *Journal of Geophysical Research*, **114**, B07203.

Main, I.G. (2000). A damage mechanics model for power-law creep and earthquake aftershock and foreshock sequences. *Geophysical Journal International*, **142**, 151 − 161.

Methods for multicausality in epidemiology: application to survival data of patients with end stage renal disease

Alicia Thiéry[1], François Séverac[2], Mickaël Schaeffer[1], Marie-Astrid Metten[1], Monica Musio[3], Erik-A. Sauleau[2]

[1] Public Health Department, Group *Methods in Clinical Research*, Strasbourg, France
[2] Laboratory of Biostatistics and Medical Informatics, *ICube*, Strasbourg, France
[3] Department of Mathematics, Cagliari, Italy

E-mail for correspondence: `alicia.thiery@outlook.fr`

Abstract: Haemodialysis and peritoneal dialysis patients survival was compared using marginal structural models or time-dependant covariates Cox model, through a large observational study. Adjustment variables were selected by directed acyclic graphs or backward procedure.

Among a total of 13,767 patients with 7,181 events (52%), 1,748 (13%) have originally a peritoneal dialysis and 19,019 (87%) have a haemodialysis. The time-dependant covariates Cox model with the DAG based method covariates selection found a death hazard ratio of 0.65 [0.59 ; 0.71] in favor of haemodialysis. Marginal structural models with same adjustment found a hazard ratio of 0.82 [0.69 ; 0.97]. Similar results were found with backward procedure.

Marginal structural models differ from Cox model in considering clinical characteristics by using weights. A benefit of the graphical method is to allow for the representation of prior clinical knowledge on a given situation (including unobserved variables) and to have an overview of the causal paths.

Keywords: Observational study; Causality; Directed acyclic graphs; Marginal structural model; Survival

1 Introduction

The choice of dialysis modality, haemodialysis (HD) or peritoneal dialysis (PD), has become an important decision that affects patients quality of life and survival. There are conflicting research results about the survival

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

differences between HD and PD. Because a lot of factors act together on causal relations, epidemiologists face very complex situations and describe causes of disease in a multifactorial framework. Much of the data used for causal claims do not come from experiments such as clinical trials and one of the most important issue is then whether it is possible to make warranted causal claims using non-experimental, observational data. This requires an examination of "causal criteria" but a more sophisticated reasoning about causality (an event C causes an event E) is also needed.

We are in a situation of time-dependant covariates (patients can change of dialysis modality), resulting confounding bias. Appropriate Cox model could be used but in this example it seems that inverse probability weighting and marginal structural models are better analytic tools to be used to avoid the bias that can occur with standard adjustment of a time varying confounder affected by prior exposure. Then still remains the problem of adjustment covariates choice. The most used methods are stepwise procedures but because they may introduce biases by themselves, like collider bias or confounding bias (Greenland, 2003), alternative methods have been suggested, as directed acyclic graphs (DAGs).

The aim of the study is to compare the survival of patients in haemodialysis and patients in peritoneal dialysis in a large national cohort of incident dialysis patients. We used a directed acyclic graph to select adjustment variables and marginal structural models to account for transplant censorship, modality change over time and time varying covariates.

2 Materials and methods

Data came from REIN (*Réseau Épidémiologie et Information en Néphrologie*), a national register of patients with end stage renal disease in France. Pathological conditions (associated heart disease, diabetes, ...), laboratory tests and medications are annually collected. For this survival study, we selected patients who started a dialysis between 2006 and 2008 with no emergency and gathered their annual follow up until December 31, 2013.

For dealing with multicausality, several models have been suggested. Marginal structural models (MSM, Robins *et al*, 2000) rely on the counterfactual approach. Causality is then defined by comparing the observed event and the counterfactual event that would have been observed if contrary to the fact, the subject had received a different exposure than the one he actually received. MSM have been also developed to address the issue of time varying confounding (Hernan *et al*, 2000), using the inverse probability weights (IPWs). IPWs are estimated by combining the inverse probability of "treatment" (here dialysis modality) weights (IPTWs) and the inverse probability of censoring weights (IPCWs). The IPTW (or IPCW) are computed as the ratio of the estimated probabilities of treatment (or censoring) using baseline covariates only and of the estimated probabilities of treatment (or censoring) using baseline and time-dependant covariates. For example, at time t, the IPTW for subject i is $\prod_{s\leqslant t} \dfrac{P\left(T_s|\mathbf{E}\right)}{P\left(T_s|\mathbf{E},\mathbf{L}_{s-1}\right)}$, where (without a strong specification) T_s is the treatment at time s, $\mathbf{E}$ are baseline covariates and $\mathbf{L}_{s-1}$ are time-dependant covariates, observed at the time $s-1$.

Because it is possible to represent causal relation by graph (two vertices and an oriented edge), DAGs are visual representations of the pre-supposed causal relationships between variables, including exposure covariates, outcome, potential confounding variables and also latent variables. Shrier and Platt (2008) described a set of rules based on the created DAG to decide which variables must be controlled in the statistical analysis in order to remove confounding.

With respect to our main goal, we compared four models on survival data, where the treatment under interest was the dialysis modality and an additional censor event was kidney transplant. The models compared, all dealing with time-dependant covariates, were Cox regressions with or without the use of IPWs, using for selection of covariates a backward procedure or a DAG-based method.

3 Results

Among a total of 13,767 patients, 1,748 or 13% had originally a PD and 19,019, or 87%, had a HD. A quarter of PD patients switched to HD whereas 1% switched from HD to PD. At baseline, patients in HD and PD differed on several characteristics (age, diabetes, heart failure . . .).

Five adjustment covariates were selected with a backward procedure: age, cancer, estimated glomerular filtration rate, heart failure and chronic respiratory failure. With the DAG-based method 7 more covariates were added: transplant waiting list, stroke, kind of nephropathy, smoker, peripheral vascular disease, handicap, cirrhosis. Results of four models using or not weights and graphical method are summarized on Table 1.

TABLE 1. Estimated hazard ratio (HR) for dialysis modality (PD as reference)

Models	Adjustment method	HR [CI95%]	AIC*
Cox model	Backward procedure	0.64 [0.59 ; 0.70]	61,974
	DAG-based method	0.65 [0.59 ; 0.71]	61,463
MSM	Backward procedure	0.80 [0.68 ; 0.94]	65,244
	DAG-based method	0.82 [0.69 ; 0.97]	64,682

*Akaïke Information Criterion

4 Discussion

In our study the four models lead to the same conclusion: haemodialysis was associated with a better survival. The use of MSM with inverse probability weighting reduces nevertheless the findings of the time-dependant survival analysis. The proportional hazard ratio was 0.2 point higher, which is clinically non-negligible. This result is connected with the fact that marginal structural models consider clinical characteristics of dialysis patients across weights.

Although both methods of covariates selection gave similar estimates of hazard ratio, the benefit of the graphical method is to allow for the representation of prior clinical knowledge on a given situation (including unobserved variables) and to have an overview of the causal paths clinically defined. However graphical method increased the number of covariates used for adjustment on confounders, which can be an issue in large cohort studies using data from registers that do not always gather all the needed variables.

In addition, to avoid the problem of competing risks due to renal transplantation, we considered a second source of weights which is the inverse probability of censoring, as well as the inverse probability of treatment. A future work will be the comparison with multi-states models which also deal with competing risks. The study of models including interactions between main effects could also be explored, in order to take into account different treatment effects across subgroups (e.g. age).

References

Greenland S. (2003). Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*, **14**(3), 300 – 306.

Hernan M.A., Brumback B. and Robins J.M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, **11**(5), 561 – 570.

Robins, J.M., Hernan M.A. and Brumack B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, **11**(5), 550 – 560.

Shrier I. and Platt R.W. (2008). Reducing bias through directed acyclic graphs. *BMC Medical Research Methodology*, **8**, 70.

Partial Ranking Analysis

Tita Vanichbuncha[1], Martin S. Ridout[1], Diogo Veríssimo[2]

[1] University of Kent, UK
[2] Georgia State University, USA

E-mail for correspondence: tv64@kent.ac.uk

Abstract: Partial ranking data arise when the number of items is too large for the rankers to rank all the items. Several models for this kind of data have been proposed, including the Plackett-Luce (PL) model. An extensions of the PL model, termed the Rank-Ordered Logit (ROL) model allows covariates to be incorporated. A different extensions of the PL model, the Benter model, allows preferences for higher-ranked items to be stronger than preferences for lower-ranked items. Here we combine these two extension of the PL model to give a model that incorporates covariates and also allows for a dampening effect. We adapt the minorization-maximization (MM) algorithm that was proposed by Gormley and Murphy (2008) for fitting the Benter model.

Keywords: Partial Ranking; Plackett-Luce Model; Rank-ordered Logit Model; Benter Model.

1 Introduction

Ranking data, in which rankers are asked to express their preferences among a set of items, arise in many fields (Alvo 2014). In the simplest case, each ranker is asked to rank each item, leading to a *complete ranking*. When the number of items is too large for rankers to provide a reliable complete ranking, they can instead be asked to rank a subset of the items. This is called a *partial ranking*.

Several models for partial ranking data have been proposed in the literature (Marden 1995, Alvo 2014). One model for complete rankings that adapts readily to partial ranking data is the Plackett-Luce (PL) model (Marden 1995). The Rank-Ordered Logit (ROL) model is an extension of the PL model which can incorporate covariates (Alvo and Yu 2014). A different type of extension of the PL model is the Benter model (Benter 1994). The

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

Benter model introduces additional parameters, termed damping parameters, into the basic PL model, which allow preferences for higher-ranked items to be stronger than preferences for lower-ranked items. In our work, the two extensions of the PL model are combined to give a model that incorporates covariates and also allows for a dampening effect.
To illustrate the methodology, we use a set of partial rankings of pictures of animal species. In a survey carried out at the Durrell Institute for Conservation Ecology at the University of Kent, participants were asked to rank random samples of 10 species from a total of approximately 100 species, according to how appealing they were.

2 Models for Ranking Data

Suppose there are k items to be ranked and n individuals who are doing the ranking, where k may be large so that not all items are ranked by all rankers. Let p_i denote the number of items ranked by ranker i and let $\rho_i = (\rho_{i1}, \ldots, \rho_{ip_i})$ be the ranking where ρ_{i1} is the item ranked first, ρ_{i2} is the item ranked second, etc.

2.1 Plackett-Luce Model

The Plackett-Luce (PL) model (Plackett 1975) is a very popular parametric model of ranking data. The items are ranked from best to worst and the rankings by different rankers are assumed independent. The PL model has a vector of parameters $(\lambda_1, \ldots, \lambda_k)$, where λ_j is a measure of the preference for item j. Writing $\mu_{\rho_{ij}} = \exp\left(\lambda_{\rho_{ij}}\right)$, the PL model specifies the probability of the ranking ρ_i as

$$P(\rho_i; \lambda) = \frac{\mu_{\rho_{i1}}}{\mu_{\rho_{i1}} + \cdots + \mu_{\rho_{ip_i}}} \times \cdots \times \frac{\mu_{\rho_{ip_{i-1}}}}{\mu_{\rho_{ip_{i-1}}} + \cdots + \mu_{\rho_{ip_i}}} \times \frac{\mu_{\rho_{ip_i}}}{\mu_{\rho_{ip_i}}}.$$

2.2 Rank-Ordered Logit model

The Rank-Ordered Logit (ROL) model is an extension of the PL model that can incorporate covariates. There are three kinds of covariates describing item characteristics, ranker characteristics, and ranker-item characteristics (Alvo and Yu 2014). Let $\mu_{\rho_{ij}}$ be the corresponding ROL parameter then the general form of the model is

$$\mu_{\rho_{ij}} = \exp\left(\lambda_{\rho_{ij}} + \sum_{l=1}^{L} \beta_l z_{l,\rho_{ij}} + \sum_{r=1}^{R} \gamma_{r,\rho_{ij}} x_{r,i} + \sum_{q=1}^{Q} \theta_q w_{q,i\rho_{ij}}\right),$$

where $z_{l,\rho_{ij}}$ is a covariate that depends on the item ρ_{ij} and β_l is an item-specific parameter; $x_{r,i}$ is a covariate relating to the rankers but does not vary over items and the coefficient $\gamma_{r,\rho_{ij}}$ is a ranker-specific parameter; and finally, $w_{q,i\rho_{ij}}$ is a covariate that describes a relation between item ρ_{ij} and ranker i and θ_q is a ranker-item specific parameter.

2.3 Benter Model

Benter (1994) introduced another extension of the PL model by adding a set of damping parameters (α_j). In the Benter model, the probability of the ranking ρ_i is

$$P(\rho_i;\lambda) = \frac{\mu_{\rho_{i1}}^{\alpha_1}}{\mu_{\rho_{i1}}^{\alpha_1} + \cdots + \mu_{\rho_{ip_i}}^{\alpha_1}} \times \cdots \times \frac{\mu_{\rho_{ip_{i-1}}}^{\alpha_{p_i-1}}}{\mu_{\rho_{ip_{i-1}}}^{\alpha_{p_i-1}} + \mu_{\rho_{ip_i}}^{\alpha_{p_i-1}}} \times \frac{\mu_{\rho_{ip_i}}^{\alpha_{p_i}}}{\mu_{\rho_{ip_i}}^{\alpha_{p_i}}},$$

where $\mu_{\rho_{ij}} = \exp\left(\lambda_{\rho_{ij}}\right)$. The Benter model is characterized by the parameters α_j and assumes that $0 \leq \alpha_j \leq 1$ for all $j = 1, \ldots, p_i$. This ensures that preferences for lower ranked items are at least as random as higher preference ones. In order to avoid over-parametrization problems, α_1 and α_{p_i} are defined to be equal to 1 and 0, respectively.

3 Statistical Analysis

Our investigation is motivated by a real data set coming from an internet survey to assess the appeal of pictures of different animal species. This survey divided the species into four groups. We consider only one group which contains 97 species of animal and there are 450 participants. We included ranker-item-specific covariates (Familiarity and Start Position) and ranker-specific covariate (Gender) into ROL and Benter models. The models were fitted by maximum likelihood and likelihood ratio tests were applied to compare models when adding one covariate at a time.

TABLE 1. Likelihood ratio statistics when adding Familiarity, Start Position, and Gender to ROL and Benter models (p-value in brackets).

Covariate	**ROL**	**Benter**	**ROL vs Benter**
None	-	-	109.43(0.00)
+ Familiarity	91.81(0.00)	112.91(0.00)	130.53(0.00)
+ Start Position	19.43(0.00)	13.18(0.00)	124.28(0.00)
+ Gender	118.00(0.06)	100.25(0.36)	106.41(0.00)

Table 1 shows that the Familiarity and Start Position are strongly significant, but Gender is not significant at 5% level when Familiarity and Start Position are already in the model. Moreover, the Benter model performs better than the ROL model which means that adding the dampening parameters to the model can improve performance.

The estimated α parameters are shown in Figure 1. The parameters are generally decreasing with rank, suggesting that the rankers ranked their top preferences more carefully their lower preferences.

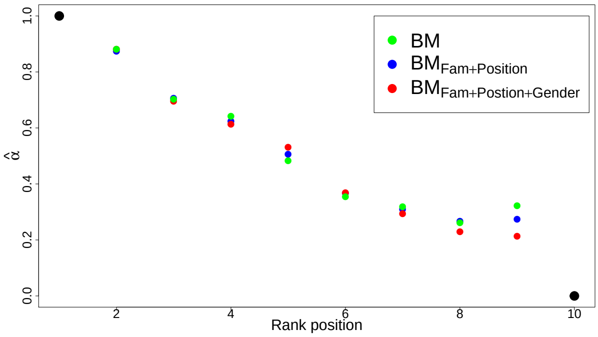


FIGURE 1. Comparing the α parameters from the Benter model.

4 Conclusion

We find that both extensions of the PL model result in significant improvements in fit to the observed partial rankings. In the extended model, the most significant covariates relate to the order in which the species were presented to the ranker and whether the species was familiar to the ranker. The damping parameters of the Benter model indicate that preferences become less strong as one moves down the ranking order.

References

Alvo, M. and Yu, P. (2014). *Statistical Methods for Ranking Data.* Springer.

Benter, W. (1994). Computer-Based Horse Race Handicapping and Wagering Systems: A Report. *Efficiency of Racetrack Betting Markets*, 183 – 198.

Gormley, I.C. and Murphy, T.B. (2008). Exploring Voting Blocs Within the Irish Electorate: A Mixture Modelling Approach. *Journal of the American Statistical Association*, **103**, 1014 – 1027.

Marden, J.I. (1995). *Analyzing and Modeling Rank Data.* London: Chapman & Hall.

Plackett, R.L. (1975). The Analysis of Permutations. *Journal of the Royal Statistics Society. Series C (Applied Statistics)*, **24**, 193 – 202.

A non-iterative approach to ordinal log-linear models: investigation of log D in drug discovery

S. Zafar[1], I. L. Hudson[1], E. J. Beh[1], S. A. Hudson[2], A. D. Abell[3]

[1] School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, Australia

[2] School of Chemistry, Australian National University, Canberra, Australia

[3] School of Physics and Chemistry, Adelaide University, Adelaide, South Australia

E-mail for correspondence: sidra.zafar@uon.edu.au

Abstract: We investigate drug-likeness by considering the relationships between surrogate measures of drug-likeness (solubility, permeability) and structural properties (lipophilicity (log P), molecular weight (MW)). We study the pair-wise association between categorised variants of the traditional parameters of Lipinskis rule of five (Ro5), namely MW and log P, and an additional parameter, polar surface area (PSA), introduced by Veber et al., (2002) across strata, where strata are defined by a molecule's druggable versus non-druggable (Ro5 compliant vs violation) status. Zafar, et al. (2013) earlier showed that logP's association with MW changed sign from significantly negative to positive for nondruggable vs druggable strata, becoming lower (positive) for nondruggable vs druggable. These findings support recent criticisms about using log P (Bhal et al., 2007). This study explores further the pairwise relationships of log P in comparison with log D, a distribution coefficient, and shows that log D does not swap sign nor magnitude in its relationship with MW; thereby it is a better lipophilicity measure. We use the Beh-Davy non-iterative (BDNI) direct estimation approach (Beh and Davy, 2004; Zafar et al., 2015) to estimate the linear-by-linear association of the pairwise relationships, within the framework of well-known ordinal log-linear models (OLLMs). We also provide correspondence analysis (CA) plots for comparing the pairwise associations of logP and log D with MW.

Keywords: Linear-by-linear association, ordinal log-linear model, log D

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

1 Introduction

Assessing drug-likeness depends on the nature of relationships between surrogate measures of physiochemical properties (aqueous solubility, permeability) and structural properties (lipophilicity, molecular weight (MW)). Lipophilicity, quantified as the partition coefficient, log P, is the most popular predictor for permeation as it influences drug potency and absorption, distribution, metabolism and excretion (ADME) properties. Lipophicility is also described by log D, a distribution coefficient, describing the ratio of sum of concentrations of all compound forms (ionised and un-ionised). Bhal et al. (2007) suggested that log P often fails to take into account variation in lipophilicity of a drug and proposed log D, as a better lipophilicity descriptor. Zafar, et al. (2013) recently showed that log Ps association with MW, optimally assumed positive, changed sign from significantly negative to positive for nondruggable vs druggable strata, and became lower (positive) for nondruggable vs druggable molecules. In this study, we test Lipinskis rule of five (Ro5) (Lipinski and Hopkins, 2004). Our aim is to conduct a comparison of log P and log D, that is, to identify whether differences in magnitude or sign change of the pairwise associations occurs across strata (strata defined by a molecules druggable (Ro5 compliant) versus non-druggable (Ro5 violation) status) by using the non-iterative direct estimation approach (Beh and Davy, 2004) to estimate the association for pairwise relationships within the framework of the ordinal log-linear models (OLLMs).

2 Mathematical methods

For a doubly ordered $I \times J$ table, N, denote the proportion in the (i, j)th cell as $p_{ij} = n_{ij}/n$ where n_{ij} is $(i, j)\,th$ cell value of N, for $i = 1, 2, \ldots.I$, and $j = 1, 2, \ldots..J$. Denote $p_{i.}$ and $p_{.j}$ as fixed marginal proportion of ith row and jth column, respectively. Let m_{ij} be the expected cell frequency of (i, j)th cell. The OLLM is then defined as

$$\ln m_{ij} = \mu + \alpha_i + \beta_j + \phi(u_i - \overline{u})(v_j - \overline{v}). \tag{1}$$

Here, u_i and v_j are row and column scores, respectively. Set $u_i = i$ and $v_j = j$. μ is grand mean, α_i and β_j are ith row and jth column effects, respectively. The parameter of interest in this model is the linear–by-linear association parameter, ϕ. The BDNI estimator is

$$\hat{\phi}_{BDNI} = \frac{1}{\sigma_I^2 \sigma_J^2} \sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} \, (u_i - \overline{u}) \, (v_j - \overline{v}). \tag{2}$$

Zafar et al. (2015) recently explored the bias, consistency, variance and relative efficiency of the BDNI estimator by comparing to the Cramer-Rao lower bound and showed that the BDNI estimator is a Minimum Variance Unbiased (MVU) estimator of the linear-by-linear association parameter, under independence. For visual interpretation of the association between ordered row and ordered column categories, a CA plot is constructed.

3 Data and Design

In this study the data of Hudson et al. (2012), namely, 1,279 small molecules from the DrugBank3.0 database (Knox et al., 2011) are analysed. We evaluate the relationships between bivariate pairs of categorised variants of 2 of the 4 traditional parameters of Ro5, namely MW and logP, an additional parameter, polar surface area (PSA) (Veber et al., 2002). Four levels of categorized variables are generated by making quartiles within the given druggability stratum (levels 1, 2, 3) and the first category (0) is defined to satisfy the new molecular cutpoints (criteria) determined by Hudson et al.,(2012).

4 Results and Conclusion

Table 1 gives BDNI estimates (95 % CIs) and corresponding linear-by-linear CA estimates for the Ro5 based druggability strata. All BDNI estimates are highly significant. In summary, log Ps association with MW changes magnitude [significant different at 5% level of significance] for the non-druggables. Log Ds association with MW does not swap sign [insignificant difference at 5%] for the Ro5 violators. The association between PSA and MW is significantly positive for nondruggable vs druggables, respectively. For demonstration, the pairwise association of logP vs. MW is reflected in ordinal CA plots (Figure 1and 2).

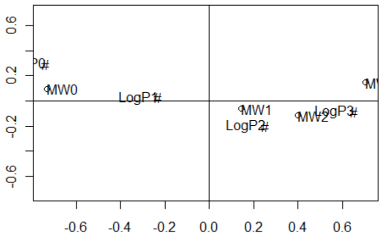


FIGURE 1. LogP vs. MW (Druggable)

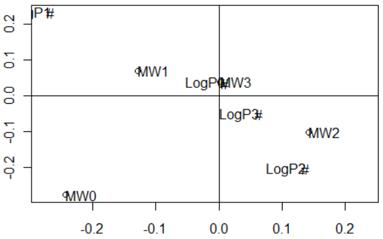


FIGURE 2. LogP vs. MW (Nondruggable)

Therefore, replacement of log D to measure lipophicility may reduce the number of false-negatives incorrectly eliminated in drug-screening. Drug-like responses are intrinsic properties of molecules and it is the responsibility of medicinal chemists to optimise molecular pharmacological properties and also drug-like properties.

TABLE 1. BDNI and CA estimates for Ro5 based Druggability (Quartile within Strata)

	$\hat{\phi}_{BDNI}$(Druggable) (95% C.I.)	$\hat{\phi}_{BDNI}$(Nondruggable) (95% C.I.)	Statistical difference
LogP vs. LogD	**0.00389** (-0.05003,0.05782)	**0.05514** (-0.02562,0.13589)	insignificant
LogP vs. MW	**0.45662** (0.38576,0.52748)	**0.043812** (-0.05730,0.14553)	significant
LogP vs. PSA	**-0.35759** (-0.42369,-0.29150)	**-0.34021** (-0.44590,-0.23453)	insignificant
MW vs. PSA	**0.24461** (0.17802,0.31121)	**0.37565** (0.18950,0.56180)	insignificant
LogD vs. MW	**0.21592** (0.16135,0.27048)	**0.26539** (0.17051,0.36028)	insignificant
LogD vs. PSA	**-0.33107** (-0.38907,-0.27306)	**-0.32947** (-0.42829,-0.23003)	insignificant

References

Beh, E.J. and Davy, P.J. (2004). A non-iterative alternative to ordinal log-linear models. *Journal of Applied Mathematics and Decision Sciences*,**7(2)**, 1 – 20. Oxford: Clarendon Press.

Bhal, S.K., Kassam, K., Peirson, I.G. and Pearl, GM. (2007). The rule of five revisited: applying log D in place of log P in drug-likeness filters. *Molecular Pharmaceutics*,**4(4)**, 556 – 560.

Lipinski, C. and Hopkins, A. (2004). Navigating chemical space for biology and medicine. *Nature*, **432**,855 – 861.Sire evaluation and genetic trends.

Hudson, I., Shafi, S., Lee, S., Hudson, S., Abell, A.D. (2012). Druggability in drug discovery: SOMs with a mixture discriminant approach. In: *Proceedings of the Aust Stats Conference, ASC 2012*,Adelaide, Australia, 105.

Knox C., Law V., Jewison T., **e**t al. (2011). DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Research*, **39** (Database issue):D1035-41.

Veber, D.F., Johnson, S.R.,*et al.* (2002). Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry*, **45(12)**, 2615 – 2623.

Zafar, S., Cheema, S.A., Beh, E.J. and Hudson, I.L. (2013). Linking ordinal log-linear models with Correspondence Analysis: an application to estimating drug-likeness in the drug discovery process. *In Piantadosi, J., Anderssen, R.S. and Boland J. (eds) MODSIM 2013. 20th International Congress on Modelling and Simulation Society of Australia and New Zealand*,Adelaide, Australia, **45(12)**, 1941 – 1951.ISBN: 978-0-9872143-3-1.

Zafar, S., Cheema, S.A., Beh, E.J. and Hudson, I.L. (2015). Properties of non-iterative estimators of the linear-by-linear association parameter of log-linear models for ordinal variables. *Submitted.*

Author Index

31st IWSM 2016 Sponsors

We are very grateful to the following organisations for sponsoring the 31st IWSM 2016.

- Centre Henri Lebesgue
- INSA Rennes
- Institut de Recherche Mathématique de Rennes (UMR 6625 du CNRS)
- Fondation Rennes 1
- Université Bretagne Loire
- Région Bretagne
- Rennes Métropole
- The Statistical Modelling Society
- Toyota Motor Corporation
- Leonard N. Stern School of Business, New York University
- Société Française de Statistique
- Presses Universitaires de Rennes
- CRCPress